CLASSIFICATION AND RECOGNITION OF ANCIENT KANNADA SCRIPTS FROM EPIGRAPHICAL DOCUMENT IMAGES

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY

In the Faculty of Science and Technology

By

SOUMYA A

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE

UNIVERSITY OF MYSORE, MANASAGANGOTRI

MYSURU – 570 006, INDIA

June – 2016

UNIVERSITY OF MYSORE

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE
MANASAGANGOTRI
MYSURU – 570 006, INDIA

DECLARATION

I hereby declare that the entire work embodied in this Doctoral thesis has been carried out by me at the **Department of Studies in Computer Science**, University of Mysore, Manasagangotri, Mysuru, under the supervision of **Prof. G. HEMANTHA KUMAR**. This thesis has not been submitted in part or full for the award of any diploma or degree of this or any other University.

SOUMYA A
Research Scholar,
Department of Studies in Computer Science,
University of Mysore,
Manasagangotri,
Mysuru – 570 006, INDIA.

UNIVERSITY OF MYSORE

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE
MANASAGANGOTRI
MYSURU – 570 006, INDIA

CERTIFICATE

This is to certify that **Ms. SOUMYA A** has worked under my supervision for her Ph.D. thesis entitled "**Classification and Recognition of Ancient Kannada Scripts from Epigraphical Document Images**". I also certify that the work is original and has not been submitted to any other University wholly or in part for any other degree.

Dr. G. HEMANTHA KUMAR
Professor
Research Supervisor
Department of Studies in Computer Science,
University of Mysore, Manasagangotri,
Mysuru – 570 006, INDIA.

UNIVERSITY OF MYSORE

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE
MANASAGANGOTRI
MYSURU – 570 006, INDIA

CERTIFICATE

This is to certify that **Ms. SOUMYA A** has worked at the **Department of Studies in Computer Science**, University of Mysore, Manasagangotri, Mysuru, under the supervision of **Prof. G. HEMANTHA KUMAR,** for her Ph.D. thesis entitled "**Classification and Recognition of Ancient Kannada Scripts from Epigraphical Document Images**".

Dr. G. HEMANTHA KUMAR
Professor
Research Supervisor
Department of Studies
in Computer Science,
University of Mysore,
Manasagangotri,
Mysuru – 570 006,
INDIA.

Dr. D. S. GURU
Professor
Chairman
Department of Studies
in Computer Science,
University of Mysore,
Manasagangotri,
Mysuru – 570006,
INDIA.

ACKNOWLEDGEMENTS

I thank the most gracious almighty for showering immense blessings on me. I wish to express my sincere gratitude to those who have contributed to my research work and supported me in one way or the other during this amazing journey.

First and foremost, I would like to express my profound gratitude and sincere thanks to my mentor and research supervisor, Prof. G. Hemantha Kumar, for his valuable guidance and being a constant source of inspiration throughout my Ph.D work. I am extremely thankful to him for the support he extended at all times over these years.

I would like to thank the doctoral committee members of the Department of Studies in Computer Science at University of Mysore Prof. P. Nagabhushan, Prof. Suresha and Prof. D.S. Guru, for scholarly discussions and suggestions provided.

I thank Prof. V. G. Talwar, Former Vice Chancellor, and Prof. K.S. Rangappa, Honorable Vice Chancellor, University of Mysore, Mysore for giving an opportunity to work towards doctoral degree in the esteemed university.

I wish to express my thanks to Dr. Lalitha Rangarajan, Smt. L. Hamsaveni, Dr. H. S. Nagendraswamy, Dr. B. Sharada and other faculty members for their support and valuable suggestions during research.

I wish to thank fellow researchers Mr. Manoj K, Mr. Chethana Kumara B.M., Dr. Vinay K., Dr. M. Imran, Mr. Naveena M and many others for all their help and support.

I am very much thankful to Dr. V. N. Manjunath Aradhya, S.J.C.E., Mysore, for his valuable advice and profound help in report write-up.

I would like to acknowledge the academic and technical support offered by University of Mysore, Mysore, for this Ph.D programme.

I express my gratitude to Dr. R. Gopal, Former Director, Directorate of Archaeology & Museums, Mysore, Dr. S.K. Aruni, Member Secretary, Indian

Council of Historical Records (ICHR) for providing source materials, books, photographs, and information related to epigraphy.

I am extremely grateful and thankful to Dr. K. N. Subramanya (Principal) and Dr. Shobha.G (Professor and Head Department of CSE), R.V. College of Engineering, Bengaluru, for their constant support and help throughout my work. I also thank, my colleagues - Teaching and Non-teaching of R.V. College of Engineering, Bengaluru for all their support and cooperation to pursue research at Mysore University.

I am extremely grateful and thankful to my Parents for their encouragement, inspiring and unfailing support throughout my life and also during my research work. I express my gratitude to my Brother, Sister, Brother-in-law, and Niece for their support.

I wish to express special thanks to my husband and son for their cooperation and support during these years while pursuing research work.

I extend my deepest gratitude to my co-sister and family for their hospitality and support during my stay in Mysore.

I would also like to thank other friends and relatives who have helped me directly or indirectly to complete this research work.

Soumya A

ABSTRACT

Inscriptions are one of the important sources of historical records, which are carved information on stones, pillars, walls, copper plates and other writing material in the ancient language. The study of inscriptions known as *epigraphy* is of much importance as they provide insights into past history, culture and civilizations. These historical records are degraded or damaged over-time due to bad storage conditions and preservation of this is important. The experts known as epigraphers decipher these inscriptions and translate into modern form. It is observed that the manual effort in reading inscriptions is difficult and time-consuming. So the process of digitization and *automatic recognition of historical records* is important which can be used by social scientists and archaeologists. This research work attempts to develop an automatic epigraphic character recognition system, which is challenging and is confronted with many difficulties.

The *automatic epigraphic character recognition system* is designed and implemented in 3 phases:- *Preprocessing and Segmentation, Classification of epigraphic scripts according to period* and lastly *recognition of epigraphic characters*. The system considers medium-quality epigraphic document images as input, which suffers from poor visibility due to the presence of noise, unwanted-marks and varying degradations. Hence, the input image is subjected to preprocessing whose objective is twofold:- to enhance ancient documents for better human perception, so as to enable an epigrapher to read the epigraphs and other to facilitate the automatic machine interpretation of epigraphic images. The phase *Preprocessing and Segmentation of Epigraphs* explore the available methods for enhancing epigraphical documents such as: - Spatial filtering techniques, Noise elimination and binarization. Next, the different segmentation approaches are explored on noise-free epigraphic documents to obtain sampled characters which are further subjected to classification and recognition. Relevant features are then extracted for the individual characters and stored in a database for training and later used for testing.

Next step involves the *classification of ancient epigraphs according to period* to determine the character set to be used during recognition. Epigraphical records are from varying periods and it is found that character set is different during these periods. It is vital to know the character set of a period to decipher the inscriptions pertaining to that period. Two models are presented for dating ancient epigraphical records: - Support Vector Machine (SVM) classifier based method using Zonal features and Random Forest (RF) classifier is designed for predicting a period of ancient Kannada epigraphs, using Normalized Central Moments and Zernike Moments features.

Once the period of the script is identified, the segmented epigraphic characters are recognized and transformed to modern form. Here *different models for recognition of epigraphic characters of ancient times with a combination of varying feature extraction techniques and classifiers* are explored such as:- Zernike features with SVM classifier ; Central and Zernike Moment features with RF classifier; Zone-based and Gabor features with Artificial Neural Network (ANN) approach; Fourier features with SVM, k-Nearest Neighbors (k-NN), ANN and Naive Bayes Classifier; SURF features with SVM, ANN and k-NN classifiers; and Fuzzy classifier is designed for recognition of ancient scripts using First-order and Second-order Statistical features.

Finally a *comparative study and performance analysis of different classifiers for recognition of epigraphical characters from different periods* are made which includes:- Central and Zernike Moment features with SVM, k-NN and RF classifier; Fourier features with SVM, k-NN, ANN and Naive Bayes classifier; and SURF features with SVM, k-NN, ANN and Multiple classifiers.

Overall in present research work, the input ancient epigraph is enhanced and converted into computer processable form, which is next subjected to identification of the period of the script and automated reading of the text. Thus, efforts are made to develop a Character Recognition system which reads the text of ancient epigraphic record and displays in modern form. The work finds scope in the department of Epigraphy, Ancient History and Archaeology for automated reading of ancient scripts. Thus assists epigraphers, archaeologists, and historians for digitization and further exploration of ancient historical records.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Preamble

Computer systems serve as an effective tool which can perform a number of sophisticated operations such as machine interpretation, and thus assist experts in various domains. One such application is the process of digitization and automatic recognition of historical records which can be used by social scientists and archaeologists. *Inscriptions* are one of the important sources of historical records, which are carved information on stones, pillars, walls, copper plates and other writing material in the ancient language [1, 2, 3]. They provide insights into past history, culture and civilizations. The study of inscriptions, known as *epigraphy*, is of great significance to the mankind. These historical records are degraded or damaged overtime due to bad storage conditions and preservation of this is important.

The experts known as epigraphers are responsible for reconstructing, translating, and dating such inscriptions. Modern readers find difficulty in reading these ancient epigraphs. There is a need for the automation of these kinds of tasks, as the expert epigraphers are decreasing in number, and also manual effort in performing such tasks can be exhausting and prone to errors [4]. Hence this research work attempts to develop an automatic epigraphic character recognition system, which is challenging and is confronted with many difficulties.

The automatic epigraphic character recognition system is designed and implemented in 3 phases:- *Preprocessing and Segmentation*, *Classification of epigraphic scripts* according to period and lastly *recognition of epigraphic characters*. The system considers epigraphic document images as input which suffers from poor visibility due to the presence of noise, unwanted marks and varying degradations. Hence the input image is subjected to preprocessing. Next the segmentation of noise-free epigraphic document is performed to obtain sampled characters. Relevant features are then extracted for the individual characters and stored in a database for training and later used for testing. Next step involves the classification of ancient epigraphs according to the period to determine the character set to be used for reading epigraphs. Finally the sampled characters of the input epigraph are recognized.

This chapter includes the following sections:- An introduction to Image Processing and Pattern Recognition is given in Section 1.2. Section 1.3 gives an overview on Document Image Processing. Section 1.4 covers the importance of Epigraphical documents and challenges in reading the same. An elaborative literature review is given in Section 1.5. The motivation behind the research work and the objectives are presented in Section 1.6 and 1.7 respectively. Section 1.8 lists-out the key highlights of this thesis. Lastly, Section 1.9 provides the organization of the thesis.

1.2 Image Processing and Pattern Recognition

Computer imaging involves the acquisition and processing of visual information by the computer. It involves two primary categories:- *Computer Vision* and *Image Processing*. In *Computer Vision* applications, the processed images are used by a computer. In case of *Image Processing* applications, the output images are for human interpretation. *Image processing (IP)* aims in performing some specific tasks such as:- *image enhancement, image analysis, object recognition* and *image understanding* as per the application requirements. Image enhancement transforms the input image into better perceivable form. *Image analysis* operation takes an image as an input and produces a numerical output or descriptions of the image objects. The process *feature extraction*, acquires the distinguishing traits of the object under interest. Object/Pattern recognition is the act of considering these unique characteristics of objects and identifying /categorizing objects within the image.

The research work taken up here is the automation of recognition of epigraphical scripts. The images of ancient inscriptions or epigraphic documents are the object/pattern under study. These epigraphic document images are of poor quality due to the presence of noise, unwanted marks and also degraded due to bad environmental conditions. Most of the scripts are skewed and contain lines and characters touching each other. Hence, major Preprocesssing is required to transform the input epigraphical document into a well-formed image, before the classification and recognition of epigraphs is taken up. The major goals of image processing here are two-fold:- one is to improve the quality of input epigraphic image for better interpretation by epigraphists and other is to facilitate the automatic machine interpretation of epigraphic images.

The automatic machine interpretation is a Pattern Recognition (PR) problem, with epigraphic scripts/characters as the object/pattern under study. It is viewed as a handwritten Optical Character Recognition (OCR) problem, which involves 2 steps:- identifying the period of epigraphic scripts and recognition of noise-free epigraphic documents using character bank pertaining to the era. This requires the design of suitable pattern classifier. Overall in this research work the input ancient epigraph is enhanced and converted into computer processable form, which is next subjected to automated reading of the text. Hence, this problem can be categorized under the area of Image Processing and Pattern Recognition which involves the steps image enhancement, image analysis and object/pattern classification.

1.3 Document Image Processing - An Overview

Traditionally, paper is the medium of a document presented using ink, either by handwritten or printed means. Through time, documents have also been written with ink on palm leaves, carved on stone or engraved on copper plates. Vast amounts of historical handwritten texts are available in libraries. Digitization of these documents helps in the preservation of these age-old records. Thus, it gives a path to the objective of dealing with the flow of electronic documents in an efficient and an integrated way. The final goal is to use computers for reading and understanding the documents as humans do.

Creation of good quality digitized documents for better human perception and further for comprehension, is the goal of the field *document image processing*. The document image processing involves text processing and graphics processing. The document image processing involves three basic steps, namely *Document Image Analysis, Document Image Recognition and Document Image Understanding* [5, 6].

1.3.1 Document Image Analysis

Document image analysis [6] involves both the physical decomposition of a page and the derivation of the logical meaning or semantics of the salient fields or regions defined in the image. In general, the analysis involves the extraction and use of attributes and structure relationships in the document in order to label its components based on document type. In addition to this, the text processing deals with the text components of a document image and the tasks include recognition of the text, determination of skew angle and finding paragraphs, text lines and words.

1.3.2 Document Image Recognition

The next step is to analyze a document and segregate the text block, graphics block, picture block, etc, so as to facilitate the labeling of the blocks. The process of labeling the blocks is called document image recognition or identification.

1.3.3 Document Image Understanding

Document image understanding involves recognizing the characters on a page and assembling them into the format suitable for text processing [6]. Automatic image interpretation is often desirable as it is very fast, robust, flexible, reliable and very accurate in performing sophisticated operations compared to human/manual interpretation. Machine recognition of characters involves the computer receiving input from different sources, processing and recognition. Handwritten Character recognition can be divided into two categories namely:- Offline and Online methods [6]. On-line handwriting recognizers normally use Personal Digital Assistance (PDA), or Tablet PCs, and their performances are fairly acceptable for processing handwritten letters and symbols. In contrast, off-line systems are equipped with scanners and printers.

The generic steps in off-line Handwritten Character Recognition system as indicated in Figure 1.1 consist of:- Preprocesssing, Segmentation and Recognition.

The Document Analysis and Recognition techniques find application in many domains and some are listed below:

- Document Processing in sectors such as:- banks, post-offices, business, web based applications.
- Digital libraries - for improved access, processing, information extraction and the analysis of digital documents.

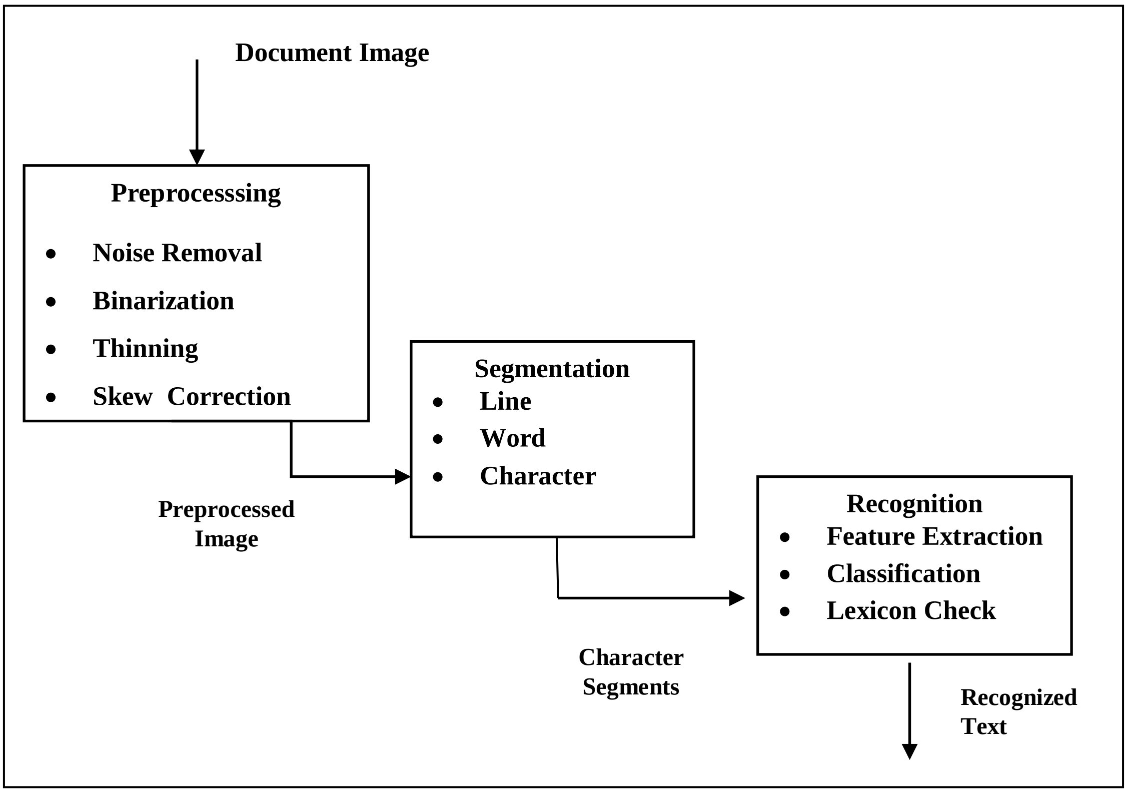


Figure 1.1: General Steps in Handwritten Character Recognition System

In this research, the problem of automatic recognition of ancient documents is addressed which reveals information of the past. Processing of handwritten historical records is important, so that it can be easily accessed and preserved for later use. The problem can be categorized under the domain of document analysis and recognition.

1.4 Epigraphical Documents and its Challenges

Epigraphy is the study which deals with the deciphering of ancient inscriptions on rocks, pillars, plates and other writing material. It is the main source for the reconstruction of history and culture, also useful to know the languages, scripts, religious and social life during the early period [1, 2, 3]. These inscriptions are found to be having broken characters, erased characters and unwanted marks. The epigraphical records are non-uniform in their sizes and non-linear in their shapes. The poor visibility and noise, the spacing between characters and also between the lines and the skew as indicated in Figure 1.2 could complicate the process of deciphering these inscriptions.

Epigraphical records are from varying periods and it is found that character set is different during these periods [1, 2]. The evolution of Indian scripts from 3rd Century BC to 18th Century AD is depicted in Figure 1.3. It is vital to know the character set of a period to decipher the inscriptions pertaining to that period.

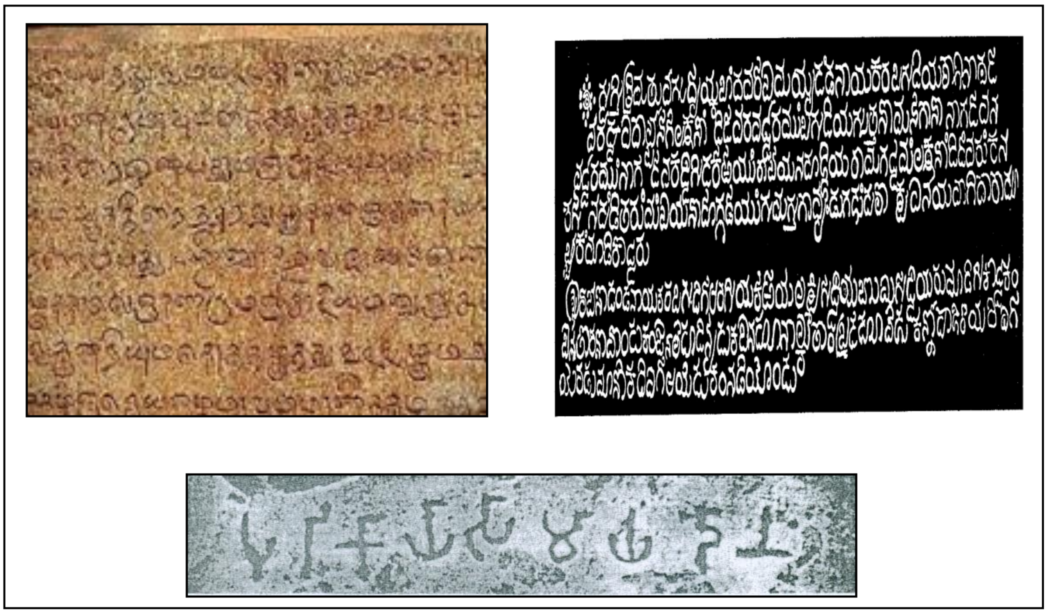


Figure 1.2: Images of Sample Inscriptions

(Courtesy: Indian Council of Historical Records, Bengaluru)

The experts called epigraphists decipher this information written using the characters pertaining to a particular era and translate these inscriptions into regional languages. The reading and translation of the ancient epigraphs are difficult and time-consuming process if done manually by epigraphists. The expert epigraphists are few and they could become extinct in future. Modern readers find difficulty in reading these ancient records. Hence, there is a need to develop an automatic system to decipher ancient inscriptions [4].

The automatic processing of epigraphical records and recognition is challenging and is confronted with many difficulties due to the storage condition and the complexity of their content [8].

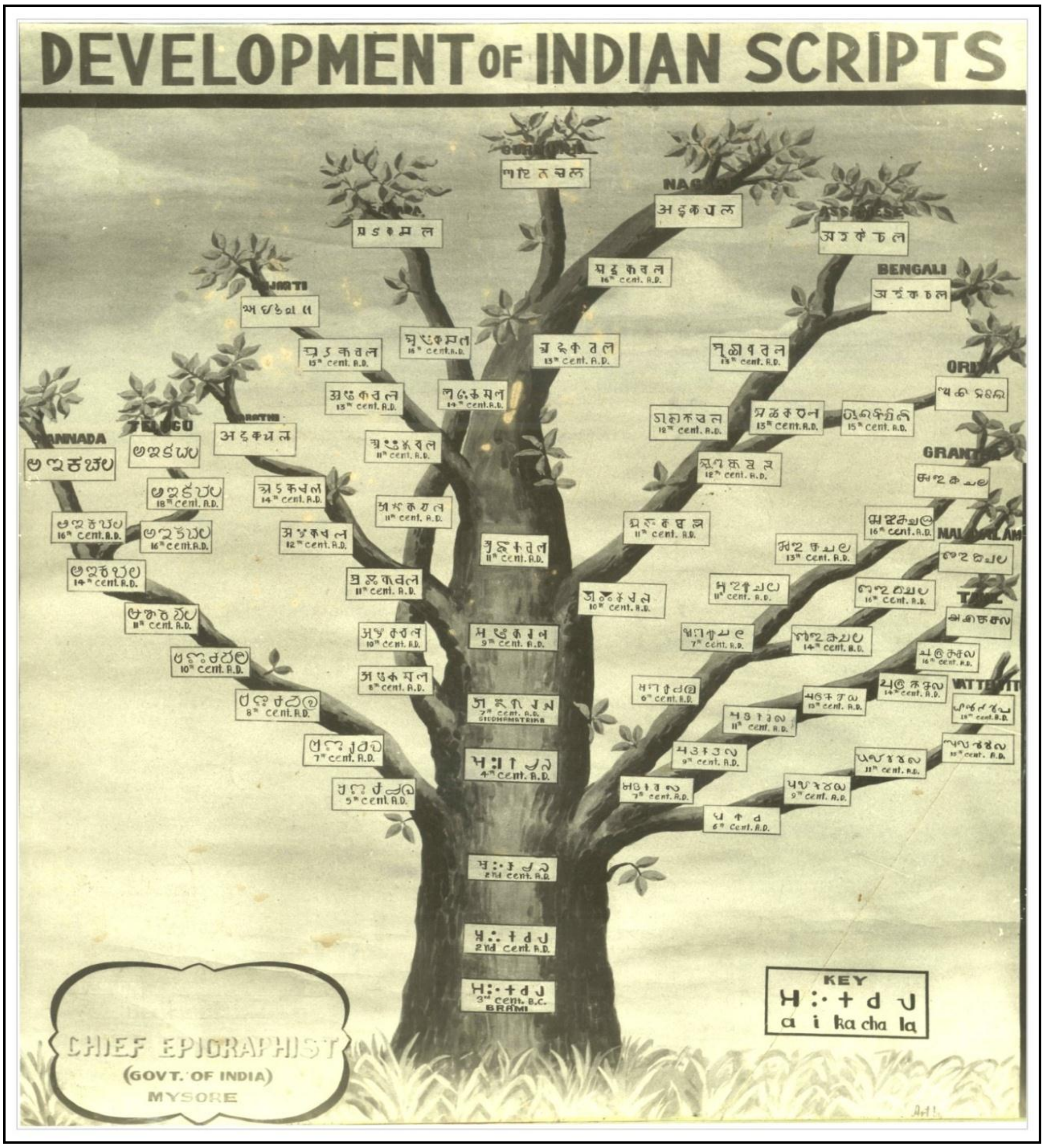


Figure 1.3: Development of Indian Scripts Since 3rd Century B.C.

(Courtesy: Dept. of Epigraphy, Archaeological Survey of India, Mysore)

Some of the challenges in automatic reading of epigraphs are listed below:

- Design an automated epigraphical document recognition system with reasonable recognition accuracy, regardless of the quality of the input document and character font style variation.
- Segmentation of handwritten text of ancient epigraphs is challenging because of its structural complication, touching lines as well as characters, non-uniform spacing, and the presence of compound characters.
- The classification of the epigraphical script according to their period is very vital in determining the character set to be applied for supervisory reading.
- Greatest challenge is the lack of standard/ benchmark data corpus to aid the recognition of ancient Indian scripts.

1.5 Related Work

An elaborative survey was carried-out in the process of work. This section presents few of the related works cited in the literature. The available techniques and methodologies used in the field are studied. Thus the reviews assisted in identifying research gap and devise a suitable model to address the current problem.

1.5.1 Preprocessing of Epigraphical Document Images

Ancient epigraphic documents are of poor quality, due to noise embedded and varying amount of degradations. Several Image processing techniques are available to address these problems and transform the input epigraph into readable form. Few related works are discussed here in this section.

The problems encountered during digitization and preservation of inscriptions is perspective distortion and the minimal distinction between foreground and background. In general inscriptions neither possess standard size and shape nor color difference between the foreground and background are addressed. An NGFICA based enhancement of inscription images is proposed in [9, 10] and this method improves word and character recognition accuracies of the OCR system by 65.3% (from 10.1% to 75.4%) and 54.3% (from 32.4% to 86.7%) respectively.

A novel method based on run length count [11] is proposed to denoise the images. In this approach first the noisy image is binarized, and then the noise is eliminated using the horizontal and vertical run length count. The algorithm is tested with noisy epigraphical document images and noisy printed document images.

Restoration plays a vital role in enhancing the degraded image. A novel approach to restoration of degraded historical document image yields better results in OCR [12]. A combination of the spatial domain along with set theory is selected to enhance the historical image. It eliminates uneven background, noise, improves the script image quality and helps in preserving the history.

A novel non-uniform slant correction preprocessing technique to improve the recognition of offline ancient Tamil text lines in [13] is presented. The local slant correction is expressed as a global optimization problem of the sequence of local slant angles. This is different to conventional slant removal techniques that rely on the average slant angle. These techniques perform well under the assumption that the text line is written with a constant slant.

Binarization is the initial step for processing, with the fact of degradation of the source document, whichever global or local thresholding approaches are chosen. The Otsu thresholding algorithm using the histogram shape analysis is most widespread global binarization algorithm. Another method proposed in [14] utilizes image contrast defined as local image minimum and maximum when compared with the image gradient process. This method is superior when handling document images with difficult background variation. Finally, the ancient document image is binarized based on the local thresholds that are derived from the detected high contrast image pixels when the same is compared with the previous method based on image contrast, the proposed method uses the image contrast to recognize the text stroke boundary and it can be used to produce high accurate binarization results.

The region based local binarization algorithm for handwritten ancient documents is used in [15] for degraded documents. In [16] thinning technique for OCR of Brahmi script is proposed. Noise elimination techniques for the Epigraphical Script images are addressed in [17, 18].

1.5.2 Segmentation of Epigraphical Document Images

Segmentation is an essential component in any handwritten recognition system. Some of the challenges to be addressed are the handwriting with diverse styles, sizes, characters which are touching or overlapping and are illegible.

Nom historical documents between 10th and 12th century AD are segmented using Projection profile based methods as discussed in [19]. Segmented characters are then categorized using K-means clustering Algorithm.

The problem of segmenting the text lines of ancient Thai manuscripts written on palm leaves is addressed in [20]. The method uses an Adaptive Partial Projection (APP) technique by integrating a Modified Partial Projection (MPP) and smooth histogram with recursion. This in comparison with the MPP looking at vowel analysis and touching components of two consecutive lines suggested that the proposed approach for practical data on palm leaf manuscripts has better performance. Percentage of components segmented within correct line was 95.99%.

Another novel method for segmentation called Nearest Neighbor Clustering for the noise-free epigraphic image is proposed [21, 22]. This method segments the line and character and works even for skewed documents. In [23] Partial Eight Direction based algorithm for segmentation of lines from epigraphic images is proposed. In work [24] techniques are proposed for detection and correction of skew and also segmentation of handwritten Kannada document. Segmentation of touching characters using contour based shape decomposition is discussed in [25]

1.5.3 Feature Extraction

In this section, few works on feature extraction techniques for recognition of old documents are discussed.

Fourier and Wavelet features are used for recognition of ancient Tamil wall inscriptions [26]. The images of the Tamil characters are pre-processed using line separation and thinning. Fourier features are extracted from the pre-processed image. Using a neural network classifier a recognition accuracy of 98% is achieved. A set of invariant moments were be extracted as features from middle age Persian characters

[27] and using a k-NN classifier an accuracy of 95% for smoothed images and 90.5% for original images are obtained.

1.5.4 Classification and Recognition of Epigraphical Document Images

The classification stage is the main decision making stage of an OCR system and uses the feature vectors extracted in the feature extraction stage to identify the text segment according to preset rules.

Combining Zernike Moments with Regional features for Classification of Handwritten ancient Tamil scripts using Extreme Learning Machine is presented in [28]. The Extreme Learning Machine is trained by Zernike moments and Regional features. The performance of Extreme Learning Machine is compared with Probabilistic Neural Networks and inferred that Extreme Learning Machine gives highest accuracy rate of 95%.

In [29] the evolution of Brahmi script into Sinhala script on the basis of ancient Sri Lankan documents inscribed on stone surface is discussed. With the aid of modern techniques of computer image processing, precise alphabet fonts of early Brahmi scripts has been produced from photographic data of ancient Sri Lankan inscriptions. It has been shown that the produced fonts are available for establishing a method of automatic reading of ancient inscriptions by computers.

An approach for transcribing historical documents in [30] divides a text-line image into frames and a graph is constructed using the framed image. Dijkstra algorithm is applied later to find the line transcription. A character recognition accuracy of 79.3% is found in its experiments. RF Classifier has been used on the Persian language [31] to classify handwritten Persian characters. Loci features are used in the paper. A classification rate of up to 87% has been achieved. A description of the paleographic analysis of Jawi manuscripts is given in [32]. It also gives a comparison of the features, algorithms and results for paleography techniques in different languages. Some ways of computerizing paleography are described in [33]. It uses a sparse document coding for the representation of characters. An accuracy of 93.3% has been claimed in that. It also compares against three other methods which had been done before that.

Characterization of the Arabic and Latin ancient document images is explained in [34]. Regions of images having the same size are extracted from the heterogeneous base and fractal dimension method is used to discriminate between ancient Arabic and Latin scripts. It achieves 95.87% accuracy on the discrimination between Arabic and Latin ancient document collections.

A method for the dating of the Greek inscription’s content in [35] uses “platonic” realization of alphabet symbols for the specific inscription and various geometric characteristics for the features, and classifies the period according to some statistical criteria.

An efficient technique for multi-script identification at connected component level using the convolutional neural network is described in [36]. Suitable script identification features are automatically extracted and learned as convolution kernels from the raw input. It is tested on a dataset of ancient Greek-Latin mix script document images and an accuracy of 96.37% is achieved on a test dataset at the connected component level and improved to 98.40% by using a class majority in the left-right neighboring area.

[37] Proposes a texture-based approach for text recognition in ancient documents. It copes with the challenges such as degradation, staining, fluctuating text lines, superimposition of text etc. The approach is applied to three different manuscripts, namely to Glagolitic manuscripts of the 11th century, a Latin and a composite Latin-German manuscript, both originating from the 14th century.

A method of recognition of ligatures [38] in cursive scripts like Pashto recognizes ligatures having variations like orientation, font style, and scaling. The use of Scale Invariant Feature Transform (SIFT) descriptors is proposed in this to evaluate its effectiveness for representing Pashto ligatures. 1000 unique ligatures with 4 different sizes are tested and average recognition rate of 74% is obtained.

[39] presents an approach for the detection of elements like initials, headlines, and text regions, focused on ancient manuscripts. SIFT descriptors are used to detect the regions of interest, and the scale of the interest points is used for localization. It gives a detection rate of 57% for initials and headlines, and 74% for regular text.

Work on automated scribe identification on a Middle-English manuscript dataset belonging to the 14th-15th century has been presented in [40].

Identification of the patterns in the image and extracting its features are the finest task as it directly affects the classification process. The authors of the paper Statistical Analysis of the Indus Script Using n-grams discuss the advantage of using statistical feature extraction methodologies in feature extraction process [41]. As per the analysis statistical features provide 75% accuracy in the results.

An effective system for the classification of ancient handwritten documents according to the writing style has been reported in [42]. It employs a set of features that are extracted from the contours of the handwritten images. These features are based on the direction and curvature histograms that are extracted at a global level from local contour observations. Two writings are then compared by computing the distance between their respective histograms. An identification rate of 94% is obtained in this. RF Classifier's performance for Handwritten Digit recognition has been accounted in [43].

The Ancient document recognition process consists of two stages: training with collected character image examples and classification of new character images [44]. The proposed OCR builds fuzzy membership functions from oriented features extracted using Gabor filter banks. Results on a significant test led to a character recognition success rate of 88%.

The problem of recognizing early Christian Greek manuscripts written in lower case letters [45] is given. Based on the existence of closed cavity regions in the majority of characters and character ligatures in these scripts, a novel, segmentation-free, fast and efficient technique that assists the recognition procedure by tracing and recognizing the most frequently appearing characters or character ligatures is proposed. This method gives highly accurate results and offers great assistance to old Greek handwritten manuscript OCR.

The work in [46] on classification and age identification of different characters by a hybrid model is implemented in two phases. The first phase of the work incorporates an Artificial Neural Network for identifying the base character. The second phase consists of a Probabilistic Neural Network model designed for the identification of age pertaining to the base character. A system to identify and classify Telugu characters extracted from the palm leaves, using Decision Tree approach is brought to light in [47]. The decision tree is developed using the SEE5 algorithm, which is an improvement from the predecessor ID3 and C4.5 algorithm in [48]. The identification accuracy obtained is 93.10% using this method.

Much work is reported in the literature on recognition of modern Indian and non-Indian scripts. As seen, literature also reveals substantial work on preprocessing of ancient scripts which include tasks such as noise removal, thinning, binarization and segmentation. It is noticed that work on automated reading of ancient Indian script particularly ancient Kannada script is minimal. Hence, in this research work an attempt is made in automatic recognition of ancient Kannada epigraphical scripts.

1.6 Motivation

Automatic recognition of handwritten ancient text is important so as to enable easy access and preservation. The motivation behind the work is enumerated below:

- **Significance of Automatic Decipherment of Inscriptions**

  Inscriptions reveal the past history and glory of a province. Thousands of inscriptions are being excavated and it is observed that this key source of reconstructing history is under serious threat of being lost. It is important to preserve these historical records and help many scholars who are working in the field of epigraphy, ancient history and archeology. There are only a few well-versed epigraphists who can study inscriptions and translate into an understandable form [49]. The reading and translation of the ancient script in epigraphs is a difficult and time-consuming process if done manually by epigraphists. In general, automatic interpretation system is very fast, robust, flexible, reliable and very accurate compared to manual interpretation [4]. Hence, it would be good to have automatic recognition system which does the job of an epigrapher.

- **Challenges in Deciphering Epigraphic documents**

  Modern readers find difficulty in reading the documents of ancient times. Recognition of ancient epigraphic scripts is relatively complex because of the varying degradation in the input, the nature of the script, variations in writing style. The poor visibility, noise, skew, touching lines and characters, broken and erased characters pose problems in deciphering epigraphic records. Hence, enhancement of historical documents is vital. The character shapes have changed over the centuries and different periods have a different character set, so the knowledge of characters of a script is important to read the text.

- **Few works on Recognition of Epigraphic records**

  Currently, there are many OCR systems available for handling printed and handwritten documents of modern form, for Indian and non-Indian scripts, with reasonable levels of accuracy. There are not much reported efforts in developing OCR systems for ancient Indian scripts especially for a South Indian script like Kannada. Hence, the main aim of this research work is the development of automated system for deciphering ancient Kannada scripts.

1.7 Objectives

Historical records like inscriptions play a vital role to mankind in knowing the history and civilization of ancient periods. The main objective of this research work is to design and develop an automated system for classification and recognition of epigraphic documents from different periods. It involves the stages: preprocessing and segmentation of epigraphic documents, classification of epigraphs into the respective period and finally recognition of the epigraphic characters.

- The objective of the preprocessing step is twofold: Firstly to enhance ancient documents for better human perception, so as to enable an epigrapher to read the epigraphs. Secondly, to bring epigraphic records into computer readable format such that a noise-free epigraphic image can be further input to recognition stage.

- To sample/extract individual characters from epigraphic records applying suitable segmentation approaches, so that the sampled characters can be further subjected to classification and recognition.
- To identify the period of epigraphic record, so that appropriate character set is used for automated reading. Hence, the process of building a system for automatic reading and deciphering ancient documents is made easier, by reducing the dictionary of characters to be used for the recognition.
- Finally, once the period of the script is identified, the segmented epigraphical characters are recognized and transformed to modern form.

Thus, an Character Recognition system which reads the text of ancient epigraphic record and transform into modern form is to be developed.

1.8 Highlights of the Thesis

The rich and glorious history of a region is recorded in the form of inscriptions/epigraphic records. The department of Ancient History and Archaeology is excavating new inscriptions and need for the automatic decipherment of such inscriptions is increasing, which minimizes the work or eliminates the need of an epigrapher in deciphering ancient epigraphs.

The **key highlights** are presented below:

➢ Address the challenges in enhancement and segmentation of ancient epigraphic documents such as:- varying degradation, unwanted symbols or marks, noise embedded and text engraved with much skew.

- **Preprocessing and Segmentation of Epigraphs** explore the available methods for enhancing epigraphical documents such as:- Spatial filtering techniques and Noise elimination using Rectangle Fitting method, binarization of the enhanced image using Otsu thresholding, Brensen, Niblack, Sauvola and Savakis algorithm. Characters are subjected to thinning using Guo Hall algorithm

  - Presents the algorithmic models for segmentation of epigraphical documents: Connected component labeling, Contour based Convex Hull, Nearest Neighbor Clustering, Drop Fall and Water reservoir techniques.

- ➢ Determination of period of input epigraph helps in identifying the character bank to be used for automated reading of epigraphs.

  - **Classification of Ancient Kannada Epigraphs into different Periods**, addresses dating of Kannada epigraphical records into their respective period.

  - Two models are presented for dating ancient epigraphical records: SVM Classifier based method using Zonal Features.

  - Random Forest Classifier is designed and implemented for predicting a period of ancient Kannada epigraphs, using Normalized Central Moments and Zernike Moments features.

- ➢ Automatic recognition of epigraphical records and transform into modern Kannada form.

  - **Recognition of Epigraphical Characters** presents different models to decipher text of ancient times with a combination of varying feature extraction techniques and classifiers.

  - The methods are:- Zernike Features with SVM Classifier; Central and Zernike Moment features with RF Classifier; Zone-based and Gabor features with Neural Network approach; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier; and finally SURF features with SVM, ANN and k-NN classifiers.

  - Fuzzy Classifier is designed and implemented for recognition of ancient scripts using First-order and Second-order Statistical features.

- ➢ Demonstrate the experimental results and performance analysis of different classifiers for recognition of epigraphical records.

- **Comparative Study of Different Classifiers for Recognition of Epigraphical Records** presents approaches:- Central and Zernike Moment features with SVM, k-NN and RF Classifier; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier; and SURF Features with SVM, k-NN, ANN Classifier and Multiple Classifier.

The proposed system can be utilized in the department of Epigraphy, Ancient History and Archaeology for automated reading of ancient scripts. Thus, the work assists epigraphers, archaeologists, and historians for digitization and further exploration of ancient historical records.

1.9 Organization of the Thesis

The thesis focuses on the development of an OCR system to recognize ancient Kannada epigraphical documents. It is organized into seven chapters as follows:

Chapter 1 - Introduction gives an introduction to the research work. It presents a brief overview of the domain, epigraphy and its relevance, related work, motivation and objectives towards this research work.

Chapter 2 - Epigraphical Document Analysis and Recognition provides the state of art development in the field of epigraphical document analysis and recognition. An overview of epigraphical scripts, its sources, evolution and challenges encountered in decipherment is given.

Chapter 3 - Preprocessing and Segmentation of Epigraphs explore existing techniques of Spatial Filtering and Noise Elimination for enhancing epigraphical documents, followed by binarization. It also presents the algorithmic models for segmentation of epigraphical documents. The experimental results and performance analysis are demonstrated.

Chapter 4 - Classification of Ancient Kannada Epigraphs into different Periods presents two models for dating ancient epigraphical records:- SVM Classifier based method using Zonal Features and the Random Forest Classifier using Central and Zernike moment features. The experimental results and performance analysis of these models are illustrated.

Chapter 5 - Recognition of Epigraphical Characters and Approaches explores different models to decipher text of ancient times with a combination of varying feature extraction techniques. The experimental results and performance analysis of these approaches for recognition of epigraphical documents are presented.

Chapter 6 - A Comparative Study of Different Classifiers for Recognition of Epigraphical Records presents the experimental results and comparative performance analysis of feature extraction techniques using varying classifiers for recognition of epigraphical records.

Chapter 7 - Conclusion and Future Avenues provides the concluding remarks highlighting the major contributions towards this research work and avenues for further research in automation of deciphering of epigraphical documents.

An extensive bibliography is listed at the end.

Chapter 2

EPIGRAPHIC DOCUMENT ANALYSIS AND RECOGNITION

2.1 Premise

In the preceding chapter the importance of the study of ancient inscriptions, their scope and the need for automation of deciphering epigraphical scripts was discussed. Also, a detailed literature survey of the works carried out in the field was reported. This chapter presents an exhaustive review on Indian epigraphy, the state of art development in epigraphic studies, the evolution of Indian scripts epigraphical documents, the challenges to be addressed in the field of Epigraphical Document Analysis and Recognition. The purpose of this review is to provide background information on the in the field, determine the issues to be considered in this thesis and to emphasize the relevance of the present study.

This chapter is organized into the following sections: An introduction to Indian Epigraphy is given in Section 2.2. Section 2.3 highlights Epigraphical Studies – State of Art Development. A brief review on the evolution of Indian scripts, in particular, Kannada script, is presented in Section 2.4. An overview of Epigraphical documents is discussed in Section 2.5. Epigraphical Script Recognition and the challenges are addressed in Section 2.6. Lastly, Section 2.7 summarizes the chapter.

2.2 Introduction to Indian Epigraphy

India is one of the ancient countries in the world with its unique knowledge, culture and civilization. The most important source of information on our history after literature is inscriptions. The inscriptions are the rare wealth of our country and authoritative sources to tell us about the life of our ancestors [1, 2, 49]. Inscriptions are also useful to know the languages, scripts, tax, administration religious and social life during the early period. Epigraphy is the study of inscription and one of the main sources for the reconstruction of history and culture. Indian epigraphy becomes more widespread over the 1st millennium AD, engraved on the faces of cliffs, on pillars, on tablets of stone, drawn in caves and on rocks. Later they were also inscribed on palm leaves, coins, copper plates, and on temple walls. Thousands of inscriptions have been copied over a period of more than 100 years from different parts of India. All of them are not maintained in good condition and many of them are almost lost. Scholars say that in the southern Indian belt, Karnataka has the maximum number of inscriptions when compared to the other Indian regions [49]. Till now more than 30000 records are found in the Karnataka region. Only 30-35 % of them have been published with notes and exhaustive study. A key to reconstructing history is under serious threat of being lost. Latest reviews have revealed that the country now has no more than 30-35 well-versed epigraphists who can study inscriptions, many of whom have either retired from active service or keep ill health. Concerned about the state of affairs, a number of government organization and forums have stepped in to chart out a number of schemes to revive the study of epigraphy in the country as discussed in section 2.3.

2.3 Epigraphical Studies – State-of-the-Art Development

The inspection and scientific research of inscriptions began in India only after the arrival of British. Englishmen Dr. Fleet and B.L. Rice were instrumental in laying the foundation for this research. Benjamin Louis Rice (B.L. Rice 1837-1927) rightly known as ‘Father of Kannada Epigraphy’ has edited, translated, and transliterated thousands of inscriptions [49]. He had laid a strong basis for the study of epigraphy and archaeology through the work "Epigraphica Carnatica".

The Archaeological Survey of India (ASI) is the premier organization for the archaeological researches and protection of the cultural heritage of the nation [51]. This branch caters to the needs of the Sanskrit and Dravidian inscriptions & coins (Headquarters at Mysore) with the functions: (i) Survey, documentation, estampaging of inscriptions (both stone and copper plates); (ii) Survey, documentation of coins; (iii) Decipherment, research, study and publication of research.

The Epigraphical Society of India brings out epigraphical discoveries made in different parts of the country. The journal of the Epigraphical Society of India will provide plenty of information which can be utilized to reconstruct our past.

The Indian Council of Historical Research (ICHR) has already brought the Karnataka Volumes in the form of C.D and further plans for digitization of the source materials and the study of scripts - development of Software etc [52].

The Karnataka State Archives is a great repository wherein more than 30,000 inscriptions are available [49]. They are published in the volumes of Epigraphia Indica and Indian Antiquary, and the regular epigraphical series like Epigraphia Carnatica , South Indian Inscriptions, and Mysore Archaeological Reports.

The department of Archaeology of the various universities, museums, Kannada Sahithya Parishat, Karnataka Itihasa Academy, and many others has undertaken the study and publication of inscriptions [49].

2.4 Evolution of Indian Scripts

Palaeography is the study of ancient writing, which includes the practice of deciphering, reading, and dating historical manuscripts, and the cultural context of writing, including the methods with which writing and books were produced, and the history of scriptoria.

2.4.1 Language and Script

Indus valley, Brahmi, and Kharosti are three important varieties of scripts that were prevalent in ancient India. The scripts of modern Indian languages have evolved from one of these scripts over the centuries [1, 2, 50]. Many of the Indian scripts such as Kannada, Tamil, Telugu, and Malayalam were originated from 3rd Century Brahmi model. The evolution of the script is dependent on many factors such as the writing material, (Stone, Copper, Palm leaf, Paper etc) writing tools, modes of writing and the background of the scribes.

2.4.2 Evolution of Kannada Script and Antiquity of Kannada Language

The Kannada script has been used to write in the Kannada language which is one of the most an enriched language in India with is long historical heritage and is the official language of Southern Indian state of Karnataka [50]. Karnataka has rich historical and cultural heritage. History is well preserved here with innumerable inscriptions, and other historical artifacts and literature. Kannada script has undergone many changes across various stages and evolved to modern form.

- **The evolutionary process of the Kannada language** [1, 2, 49, 50] in inscriptions is categorized as:
  - **Poorvada Halegannada or Pre-ancient Kannada** (450 to 800 AD): The first written record in the Kannada language is found in Emperor Ashoka`s Brahmagiri decree, which dated back to as early as 230 BC
  - **Halegannada or Ancient Kannada:** 9th to 14th centuries CE.
  - **Nadugannada or Middle Kannada:** 14th to 18th Century CE. **Hosagannada or Modern Kannada:** 19th Century and also much later.

The Kannada script has evolved from Brahmi script of Ashoka period and later during the period of other dynasties as depicted in Figure 2.1 the script has undergone many changes and transformed finally into a shape used today. These changes were brought about, because of regional variations caused by writing practices such as the materials used for writing the script, the tools used for writing, modes of writing and the background of the scribes [A V Narasimha murthy 1968].

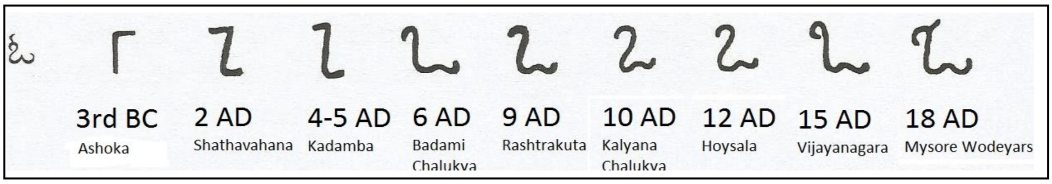


Figure 2.1: Evolution of Character ‘oo’ from 3rd B.C to 18th Century A.D

(Courtesy: Directorate of Archaeology & Museums, Mysore)

Brahmi script (3rd Century B.C): Kannada script is one among the evolved forms of Brahmi script which is found in the Ashokan edicts from places of Karnataka.

Kadamba Script (5th Century AD) **:** used by Kadambas and block hard stone were used for writing inscriptions.

Adi Ganga (4th- 6th AD)**:** used by the Gangas and resembles Kadamba Script.

Badami Chalukya (6th - 7th Century AD)**:** used by the Badami Chalukya and sand stone were used for writing.

Rastrakuta script (8th - 10th Century AD)**:** used by the Rastrakuta rulers.

Kalyana Chalukya script (10th - 12th Century AD)**:** used by the Kalyana Chalukya rulers.

Hoysala Script: used by the Hoysala Kings and is one of the most decorative forms of ancient Kannada script as soap stones were used as writing material.

Vijayanagara Script (14-16 Century AD)**:** used by Vijayanagara rulers. Most of the records are written on rough granite and hence is not decorative or uniform.

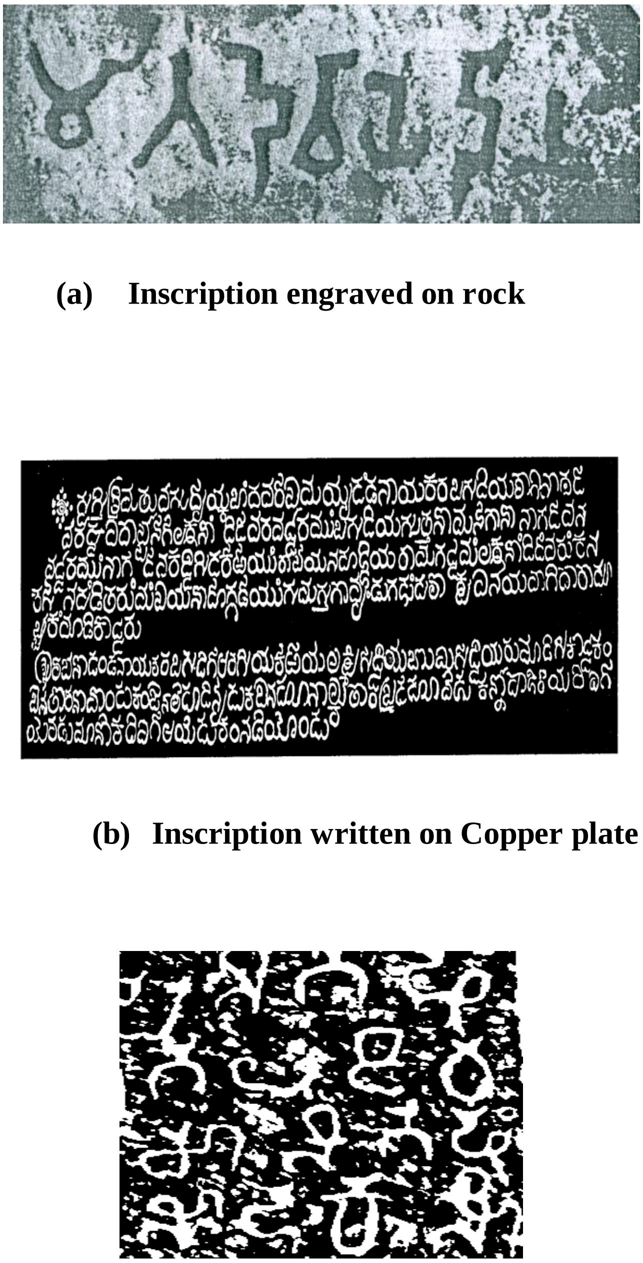
Mysore Wodeyars Script: used by Mysore kings is almost similar to the present day script and can be read easily.

This leads to the different character sets during different periods and one has to know the character set of that period to decipher the inscriptions.

2.5 Epigraphical Documents

2.5.1 An Overview

Inscriptions are source materials which throw light on the past history and civilization. Epigraphy is the study of inscriptions and the epigraphical documents were written on different materials such as rocks as in Figure 2.2(a), plates is in Figure 2.2(b) and Palmyra leaves using the characters pertaining to a particular era [1, 2, 3, 49]. Inscriptions are also available in the form of estempages - which is the imprint of the written information on white paper spread on rock as shown in Figure 2.2(c). The epigraphists translate these inscriptions into regional languages. However, these inscriptions are found to be having unwanted marks, noise, broken characters, erased characters and touching characters. Also, some estempages are worn out with time and hence a few characters have been distorted. The poor visibility and noise are also some of the facts, which are leading to problems in the deciphering of these inscriptions.



(a) Inscription engraved on rock

(b) Inscription written on Copper plate

(c) Estempage of an inscription on the rock

Figure 2.2: Sources of Inscriptions

(Courtesy: Directorate of Archaeology & Museums, Mysore)

2.5.2 Inscriptions – Sources and Types

Inscriptions vary according to the writing material. Most widely they are engraved on rocks or stones. The other sources of the inscription are palm leaves, plates made from copper, gold or silver which usually are records like donations given to institutions, sale deed, musical compositions, etc.

In general, the inscriptions engraved on rocks are:

- Edicts of the rulers: Achievements made by the rulers
- Eulogies: awards given to praise people
- Commemorative inscriptions: Donatory Inscriptions, Hero stones, Sathi stones

Inscriptions of Karnataka [49] are categorized based on the writing material as:

Shila Shasana / Stone, *Earthen Inscription* and *Loha Shasana* /Metal Inscriptions.

- **Shila Shasana/ Stone:** *Sthambha* – pillar, *Bande* – boulder, *Chappadi* – granite slab, *Karandaka* – Stone/Earthern pot
- **Earthen Inscription:** Pot / Clay or other material
- **Loha Shasana /Metal Inscriptions:** *Thamra*-copper, *Hitthale* – brass, *Kabbina* – iron, *Kanchu* – bronze, *Belli* – Silver and *Chinna* – gold.

Further inscriptions found in Karnataka are classified into the following types based on context: *Dana Shasana* – Grants, *Prashasthi* – Eulogies, *Memorial Stones*, *Gosasa* – donation of cows.

- **Dana Shasana – Grant Inscription:** are the official documents or charter conveying the grant or gift of lands by kings, officials or common man. Dana Shasanas are classified as:

  Bhumi Dana – Grant of land, *Agrahara Dana* – Gift of villages, *Pura Dana* – Gift of Town, *Umbali Dana*, *Nettaru Koduge* or Grant given for Heroic Deed, *Devalaya Dana* – Donation to the Temple, *Nirmana* – Construction of temples, lakes, wells, *Terige Dana* – Exemption of tax

- **Prashasthi Records – Eulogies:** They praise the kings and officials and are usually written by the court poet. Some of the very rare inscriptions come under this category. Such inscription gives the eulogy of a King and does not mention any grant.

- **Memorial Stones / Hero Stones:** Best of inscriptions found in Karnataka are Memorial Stones which depict a heroic story. In their commemoration the village men or his relatives erected a stone in front of the temple or in village. These memorial stones are classified accordingly to the content of their text and sculpture found on them. They are found with or without inscriptions.

  - **Veeragallu – Hero Stone:** erected in memory of heroes who sacrificed life in a battle field while defending king or public, or defending the cattle or women in distress.

  - **Masti / Maha Sati / Mastikallu – Self Sacrifice on Pyre:** Memorial stones erected in memory of a lady who invited the death by self immolation after hearing the death of her husband.

- **Darmika / Religious:** Inscriptions are also available on bells, earthen pots, etc. Label inscriptions consist of the name of the devotee or donor in temples.

2.6 Epigraphical Script Recognition and Challenges

Vast amounts of historical handwritten texts are available in libraries, where these texts are converted to digital form to preserve the information in secondary sources even if the primary sources such as ancient scrolls of text get degraded. The poor visibility, noise, skew, touching lines and characters, broken and erased characters pose problems in deciphering epigraphic records. The character shapes have changed over the centuries and different periods have different character set, so the knowledge of characters of a script is important to read the text. Also the epigraphists who translate these inscriptions into regional languages are observed to be less in number. Hence there is a need to develop an automatic system to decipher these inscriptions**.** Literature survey in Chapter 1 reveals that the epigraphic character recognition is an interesting area that is not much explored. This has motivated to take up research work on automation of deciphering epigraphical scripts. The following subsections address the problems or challenges associated in different phases such as: Preprocessing, Segmentation, Classification and Recognition of epigraphic records.

2.6.1 Preprocessing of Epigraphical Images

The present research work takes the objects with the inscriptions as the inputs. These objects may be camera grabbed or scanned if the inscriptions are engraved on a plate. The photographs of inscriptions on rock or pillars may not result in good objects. Hence estempages are produced [1, 49] and later scanning should be done to get the object image of the inscription. The images so captured still have major problems due to the facts like the broken letters, erased letters. The presence of unwanted marks engraved by the sculptor leads to the wrong diagnosis of inscriptions. Hence this requires a lot of preprocessing before the Character recognition could be taken up. The sample scanned epigraphical document shown in Figure 2.3 indicates that preprocessing methods have got to be customized suiting the requirement.

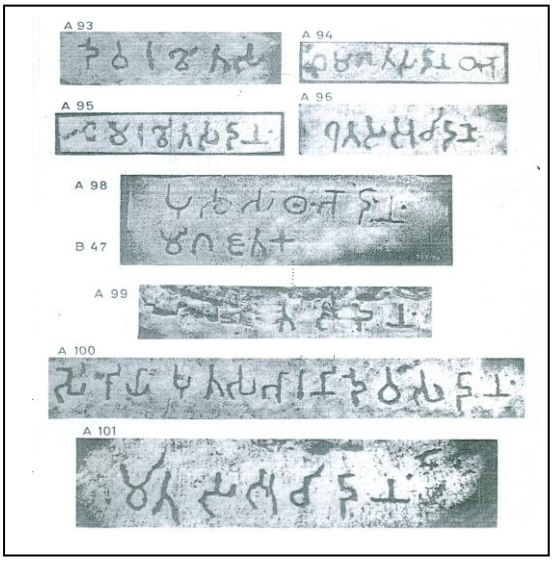


Figure 2.3: Epigraphic Records with Varying Degradation

(Courtesy: “Indian Council of Historical records (ICHR), Bengaluru)

2.6.2 Segmentation of Epigraphical Images

Segmentation of a document into lines, words and then into individual characters, constitute an important task in the optical reading of texts. Segmentation of handwritten text of Indian or non-Indian epigraphical documents is challenging because of its structural complication and presence of unwanted marks. The epigraphical records are non-linear in their shapes and non-uniform in their sizes as shown in Figure 2.4. The diverse styles and sizes of handwriting, the spacing between characters and also between the lines and the skew could complicate the process of translating the scripts. Some touching lines or touching characters or broken characters complicates the process of segmentation and could cause errors in segmentation, which result in recognition errors.

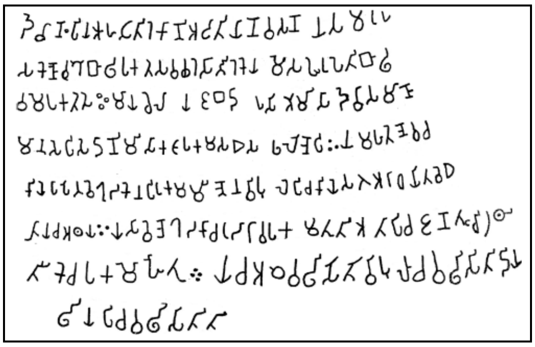


Figure 2.4: Reconstructed Epigraphic Record

(Courtesy: Directorate of Archaeology & Museums, Mysore)

2.6.3 Classification and Recognition of Epigraphical Images

In this research work, the problem of automatic recognition of epigraphical scripts is addressed which reveals details of past. Conversion of handwritten ancient text is important for making several important documents related to history, such as epigraphs, into machine editable form so that it can be easily accessed and preserved. Most of the available sources of epigraphical scripts belong to different periods, and character set varies accordingly over time as shown in Figure 2.1. The period to which the segmented character belongs is estimated and then deciphering of the complete script is taken up using the character set pertaining to that period.

2.7 Summary

This chapter covered the need and challenges to be addressed in the field of Epigraphical Document Analysis and Recognition. The complexity involved towards preprocessing, segmentation, classification and recognition of epigraphical documents are highlighted. It is noticed in Chapter 1 that work on automated reading of ancient Indian script particularly ancient Kannada script is minimal. Hence in this research work an attempt is made in automatic recognition of ancient Kannada epigraphical scripts.

Chapter 3

PREPROCESSING AND SEGMENTATION OF EPIGRAPHICAL DOCUMENTS

3.1 Premise

The details about epigraphy and its relevance, the sources of inscriptions, the need and the challenges in automation of deciphering the inscriptions are discussed in the previous chapters. The epigraphic records are of poor quality due to varying degradations, noise, and other unwanted symbols. The letters are found to be non-linear in their shapes, non-uniform in their sizes and also uneven spacing. This chapter focuses on preprocessing approaches essential for epigraphical document image analysis and segmentation of epigraphic text.

This chapter is organized into the following sections: Sections 3.2 highlights the importance of preprocessing and segmentation of epigraphical documents. The proposed model is presented in Section 3.3. The related theory and background are reported in Section 3.4. The algorithmic models for preprocessing and segmentation are presented in Section 3.5. Section 3.6 illustrates the experimental results and performance analysis of the system. Finally, Section 3.7 summarizes the chapter.

3.2 Importance of Preprocessing and Segmentation

The automatic processing of degraded historical documents is a challenge in document image analysis field which is confronted with many difficulties due to the storage condition and the complexity of the script. Raw image of an epigraph contains unwanted symbols or marks, noise embedded and text engraved with much skew. Preprocessing aims at removing the background noise and improving the readability of ancient degraded documents. Thereby these enhanced document images can be transcribed easily. This stage is followed by segmentation of characters. In epigraphic documents, the spacing between characters, also between the lines and the skew complicates the process of isolating the characters for recognition purpose. Segmentation of handwritten text of Indian languages is challenging when compared with Latin based languages because of its structural complication and presence of compound characters. This complexity increases further in recognition of ancient Indian epigraphical documents. Thus, preprocessing and segmentation plays a vital role and is challenging in the automatic decipherment of ancient epigraphs.

3.3 Proposed Model

The objective of the work is twofold: i) to enhance ancient documents with varying level of degradations and bring them in to computer readable format. ii) Followed by this to perform segmentation of ancient documents in order to extract characters from it which can be further subjected to recognition.

The proposed model for Preprocessing and Segmentation of epigraphic documents is presented in Figure 3.1. The model comprises of the components – Enhancement, Binarization and Segmentation. The input to the system is ancient epigraphic documents of varying amount of degradation. Preprocessing here mainly deals with the noise removal and enhancement of degraded ancient epigraphical images, for better human perception and also to transform the input into computer recognizable form. Enhancement is achieved through different Spatial filtering methods for smoothing or sharpening namely Mean, Median, Gaussian blur, Unsharp mask, Laplacian, and Bilateral filters. These filters are used with different mask sizes and parameter values which can be specified by the user, according to the nature of image quality and varying amount of degradation.

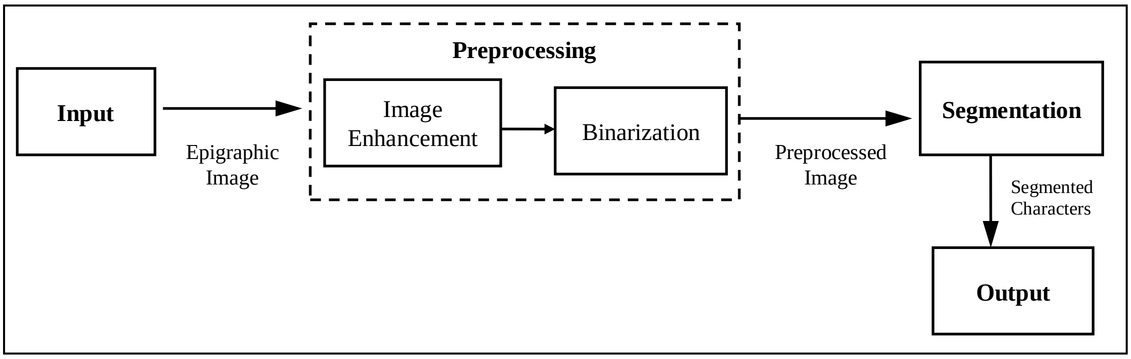


Figure 3.1: Proposed Model for Preprocessing and Segmentation of Epigraphs

Contrast enhancement is performed using Yu and Bajaj’s algorithm and binarization of the enhanced image using Otsu thresholding, Brensen, Niblack, Sauvola and Savakis algorithm. Characters are subjected to thinning using Guo Hall algorithm. Finally, segmentation of handwritten Kannada epigraphic documents is carried out using Connected component labeling, Contour based Convex Hull, Nearest Neighbor Clustering, Drop Fall and Water reservoir techniques to obtain sampled characters. The output from segmentation phase is fed to further stages of character recognition system.

3.4 The Related Theory and Background

This section covers the related theory and mathematical background of the approaches used in the proposed work.

3.4.1 Preprocessing

The purpose of extracting regions of interest and reducing noise from the images is the main purpose of the image preprocessing. The preprocessing method enhances the document image for a further analysis. Preprocessing of documents involves operations such as filtering, object boundary detection and thinning. Three classes of filter operations in DAR are noise reduction, enhancement and binarization.

- **Rectangle Fitting method for Noise Elimination**

Rectangle Fitting based noise elimination method for epigraphic characters is proposed in [17] which begins with the fixing a rectangle for each character in the image. The width and height of the characters are computed. The characters, whose width and height less than the specified width and height are considered as noisy and are eliminated.

- **Spatial Filtering techniques**

In image processing filters are designed to overturn either the high frequencies in the image which is smoothing the image or the low frequencies, which is improving or detecting edges in the image. Various filters for image processing are reported in [53, 54]. The ***mean filter*** is a sliding-window filter that replaces the center value in the window with the average of all the pixel values in that window. Thus smoothens the local variations in an image and as a result of blurring, noise is reduced. Neighborhood averaging is the best method to perform the noise reduction, whereas the method can overturn isolated out of range noise and also distorts sudden changes such as sharp edges. Under such conditions, ***Median filtering*** is widely used as it preserves edges while removing noise. In particular, this is a nonlinear digital filtering technique used to remove salt and pepper noise. It replaces the pixel value with the *median* of neighboring pixel values and suppresses the noise without damaging the sharp edges. A ***Gaussian blur*** or ***Gaussian smoothing*** is widely used to reduce the noise and enhance image structures at different scales. It is a type of image-blurring filter that uses a Gaussian function for calculating the transformation to apply to each pixel in the image. ***Bilateral filter*** is a non-linear filter in the spatial domain, which does average without smoothing the edges. Essentially the bilateral filter has weights as a product of two Gaussian filter weights, one of which corresponds to average intensity in a spatial domain, and second weight corresponds to the intensity difference. Hence, no smoothing occurs, when one of the weights is close to 0, hence preserves sharp edges.

The ***Un-Sharp Masking (USM)*** [55] technique uses a blurred or unsharp, positive image to create a mask of the original image. The unsharp mask is then combined with the negative image, creating an image that is less blurry than the original. It is a flexible and powerful way to increase sharpness, especially in scanned images. Undesired effects can be reduced, by using a mask particularly one created by edge detection to only apply sharpening to desired regions, sometimes termed smart sharpening. The ***Laplacian*** [56] is a linear derivative operator and it forms an isotropic filter. In order to get a sharpened image, typically, the resulting Laplacian filtered image is added to the original image. Negative values for the diffusion constant D will sharpen the image, positive values will blur the image.

In ***Yu and Bajaj's Contrast enhancement*** technique [57], an adaptive transfer function is designed on the basis of local statistics i.e. local maximum, minimum and average intensity and assigns a new intensity to each pixel within a local window. Using the transfer function every pixel is subjected to enhancement.

➢ Binarization Techniques

The ***Otsu's*** [58, 59] method is a popular and most efficient global thresholding technique. It is used to automatically accomplish histogram shape-based image thresholding or the reduction of a gray level image to a binary image.

Brensen's method [60] calculates local threshold value based on the mean value of the minimum and maximum intensities of pixels within a window. This threshold works properly only when the contrast is large. Also, the method is dependent on the window size and does not perform well on degraded document images with complex background.

In ***Niblack's*** method [61], the pixel-wise threshold is computed based on the local mean and local standard deviation of the rectangular window. This is an adaptive threshold method, but some noises exist in non-text regions which are to be further processed.

In ***Sauvola's*** method [62], the threshold is calculated using the local mean and standard deviation of the pixels within a window. It is a local-variance based robust method, which adapts the value of the threshold according to the contrast in the local neighborhood of the pixel. This method gives an improvement over the Niblack's method, especially when the background contains light texture, big variations, stained, badly and unevenly illuminated documents. The drawback of the method it is dependent on window size and computationally slow.

In the ***Savakis's*** algorithm [63], each pixel is assigned to a foreground or a background cluster. Pixel clustering is based on a variant of the k-means algorithm where the cluster means are updated each time a data point is assigned to a cluster. Here k=2, only two clusters are considered, which makes the overall implementation easier.

Several global and local binarization approaches for thresholding degraded ancient document images are reported in [64, 65, 66, 67].

- **Thinning**

To preserve the end points and at the same time to remove redundant pixels ***Guo-Hall*** Thinning algorithm is used in [68].

3.4.2 Segmentation

Segmentation is the method of extracting objects of interest from the image. The general steps in the segmentation of document images are identifying lines, the words in each line and the individual letterings in each word. Segmentation of text into individual letters is a major problem in recognition of handwritten documents.

- **Connected-Component Labeling and Bounding box technique**

Connected-component Labeling is an algorithmic application of graph theory, which is used the segmentation of scanned handwritten documents into characters. This approach is used in document images to segment the characters using connectivity among the components of the image.

The distribution of bounding boxes [69] describes the segmentation of an image consisting of characters. By calculating adjacency relationships merging can be performed, or their size and aspect ratios can be used for splitting mechanisms. Much of the segmentation task can be accurately performed at a low cost in computation.

- **Contour based Convex Hull and Minimum Bounding Box approach**

The approach used for segmenting the characters is a combination of Contour detection and Bounding Box Algorithm [25, 72]. First Contours are retrieved from the binary image. Once all the contours are obtained, each contour is used by Convex Hull algorithm which generates the Convex Hull objects for each character. Then the outline around each character is drawn using boundary points in a Convex Hull object. The entire image is componentized with the help of Bounding Box Algorithm. Thus, the rectangles or the bounding box are drawn across the character. To form meaningful characters or syllables some of these boxes are merged to form a single box [72].

➢ Nearest Neighbor Clustering Based Method

The majority of the epigraphical scripts contain touching lines and touching characters. Hence, a novel approach Nearest neighbor clustering method [22] for line and character segmentation is proposed, which also works even for the skewed document. A survey on methods and strategies on touched characters segmentation is covered in [73].

➢ Drop Fall Algorithm for Segmentation

Drop fall algorithm [21, 70] is based on the principle that an equally ideal cut between two touched characters can be created, if a hypothetical marble is rolled off the top of the first character and create the cut where the marble falls. The position where to drop the marble from is important because if the algorithm starts at the wrong place, the marble can simply roll down the left side of the first letter or the right side of the second letter and hence, it would be completely unsuccessful. The best approach to start drop falling process is the point at which two characters are touched. In this process, the pixels are scanned row by row until a black boundary pixel with adjacent black boundary pixel to the right of it is identified, where as the two pixels are separated by white space. This pixel is used as a point to start the drop. The direction that the algorithm will move is according to the current pixel position and its surroundings.

➢ Water Reservoir Algorithm

The larger space generated by touching numeral is analyzed with the help of water reservoir concept [71]. When water is poured from the top (bottom) of a component, the regions of the component where water will be stored are considered as a top (bottom) reservoir. The reservoirs obtained in this procedure are not considered for further processing. Those reservoirs whose heights are greater than a threshold value T1 are considered for further processing.

When two characters touch each other, they create a space (reservoir) between the characters. This space is very important for segmentation because,

- As cutting points are concentrated around the base of the reservoir, and hence, decreases the search area.
- The cutting points lie on the base of the reservoir.
- Space attributes (center of gravity and height) aid to go near the best touching position.

The touching position between characters is to be determined. The best reservoir for touching is determined. The base-line (lowermost row of the reservoir) of the best reservoir is then identified. To find the touching position in the components, the morphological thinning operation is applied to touching components for further processing. For feature points extraction the touching position is renowned. The leftmost and rightmost points of the base-line of considered reservoirs are the feature points. These points are initial feature points. With this initial feature points, the best feature point is chosen for segmentation.

3.5 Detailed Design and Algorithmic Models

The detailed design and the algorithmic models of the proposed work are presented in this section.

3.5.1 Preprocessing of Ancient Epigraphical Images

Algorithm **: Enhancement (Epigraph_Image)**

Input: Ancient epigraphical image

Functionality: Enhances epigraphic image of medium-level degradation using spatial filters namely Mean, Median, Gaussian blur, Unsharp mask, Laplacian , Bilateral filter, with varying mask size and filter parameters, depending whether smoothing or sharpening is the requirement, for better human perception. The enhanced image is subjected to binarization and thinning.

Output: The enhanced and binarized image with reduced noise.

Method:

The steps towards Enhancement of Ancient Epigraphs are as follows:

Step 1: [*Read Image*]: Read epigraphical image of varying amount of degradation

Step 2: [*Enhancement*]: The input ancient epigraph is enhanced using any of the following filtering methods, designed and implemented with different mask sizes,

```
if (method = 1 or method = 2)
    Read the size of the mask nxn , where n = 2, 3, 4, 5
    Read value of the Standard Deviation, σ
    Filter the input image using Gaussian blur method.
    Filter the input image using USM method.
Else if method = 3, (For Laplace Filter)
    Read and Display the Diffusion constant value D
    Read and display the step size of the mask, No. of Steps
    Filter the input image using Laplacian method.
Else if method =4 (For Mean Filter)
    Filter the input image using Mean filtering method
Else if method = 5(For Median Filter)
    Filter the input image using Median filtering method
Else if method = 6(For Bilateral Filter)
    Filter the input image using Bilateral filtering method
```

Step 3: [*Binarization*]: The enhanced images are converted to binary image consisting of ones and zeroes.

Use Otsu thresholding, Brensen, Niblack, Sauvola and Savakis approaches

Step 3: [*Contrast Enhancement*]: Use Yu and Bajaj's algorithm

Step 4: [*Thinning*]: Employ Guo-Hall algorithm.

3.5.2 Segmentation of Epigraphical Images

The algorithms for segmentation are presented here in the section.

➢ ***Algorithm* : Drop_Fall (Epigraph_Image)**

Input: Noise-free binarized epigraphical image

Functionality: The binarized image is segmented to characters using Drop Fall algorithm

Output: Segmented characters

[Step 1]: Find the size of the characters to find touched characters

Find the Height and Width of the touched characters

[Step 2]: Apply Breadth First Search (BFS) algorithm to find the touched characters.

Search for the initial pixel; scan row by row until two black pixels are separated by white pixel.

[Step 4]: If found start the Drop fall.

drop falling algorithm will always move downwards, crossways downwards, to the right, or to the left.

[Step 5]: Make the slice where marble parks.

Segmentation path for connected components is found

➢ ***Algorithm* : Water_Reservoir (Epigraph_Image)**

Input: Noise-free binarized epigraphical image

Functionality: Binarized image is segmented to characters using Water reservoir concept

Output: Segmented characters

[Step 1]: Find the size of the characters to find touched characters.

[Step 2]: The positions and sizes of the reservoirs are analyzed.

[Step 3]: A reservoir is detected where touching is made,

The initial feature points for segmentation are noted.

[Step 4]: The best feature points are noted from the initial feature points.

[Step 5]: Based on touching position, close loop positions and morphological structure of touching region the cutting path is produced.

3.6 Experimental Results and Analysis

This section discusses the experimental results and comparative analysis of the approaches used in preprocessing and segmentation of epigraphic records.

3.6.1 Experimental Dataset

The experimental dataset consists of degraded ancient epigraphic document images for preprocessing from sources such as camera-captured and scanned records. The subsystem for noise removal and enhancement is tested on nearly 250 epigraphical images (100 camera-grabbed images of inscriptions and 150 scanned ancient epigraphic images) with a medium amount of degradation.

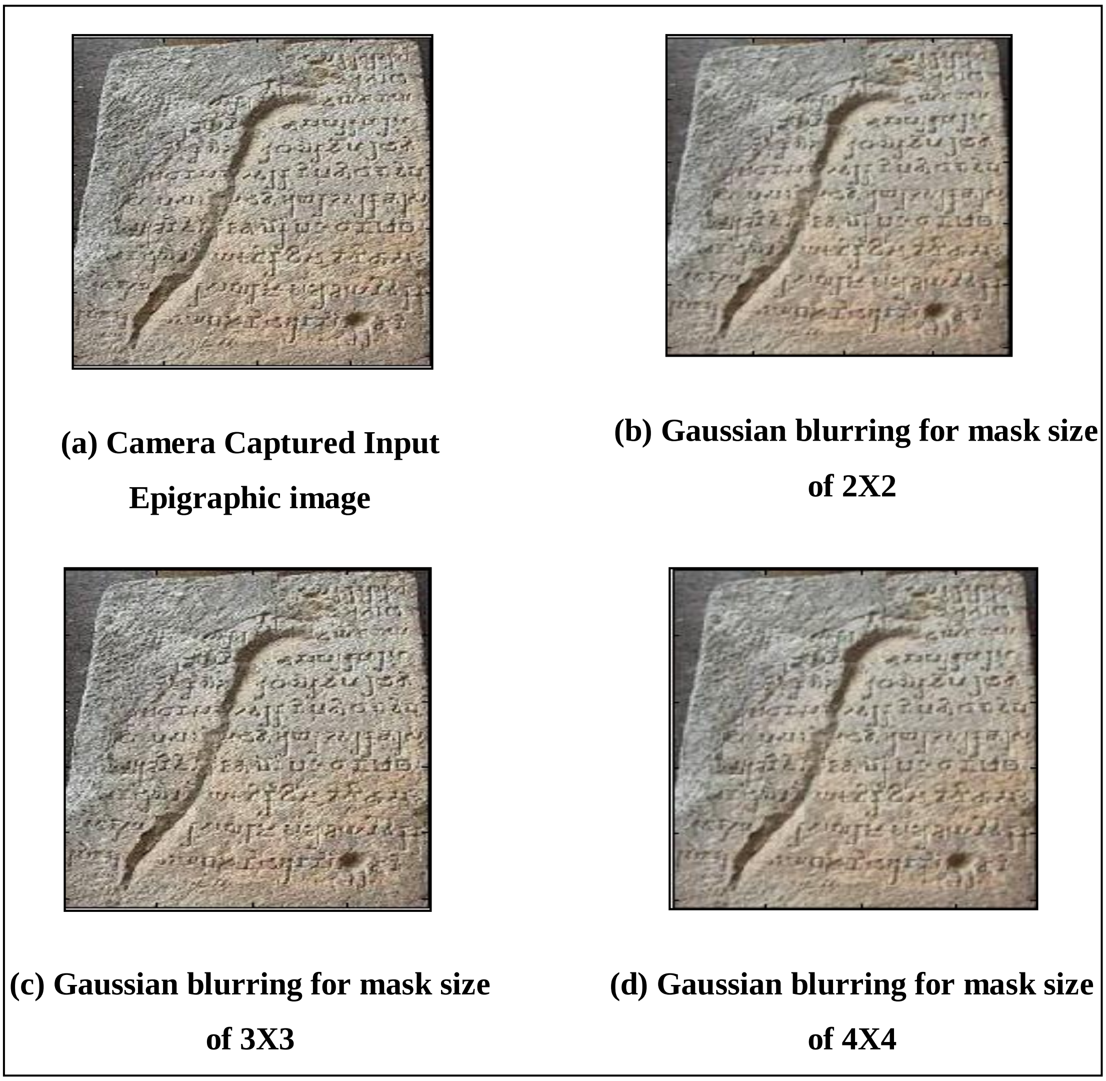
The experimental results of preprocessing and segmentation for samples of different sources, varying image quality and degradations are illustrated in the following subsections.

3.6.2 Enhancement of Camera-captured Inscriptions and Analysis

In this section, the experimental results of image enhancement using the available spatial filtering techniques such as Gaussian blur, USM, and Laplacian are explored for camera captured inscriptions with varying quality. The results of these filtering operations for different mask sizes are discussed with illustrations.

Figure 3.2 represents the results of Gaussian blur filtering, which successfully smoothens the input image and thus reduces the background noise. The Gaussian blur method is used to blur the sharpen image so that a less edge highlighted image is produced and this also reduces some amount of background noise for further stages of preprocessing and segmentation system. Figure 3.2(a) is the input epigraphic image and the Figure 3.2 (b), (c), and (d) shows various results of Gaussian blur method for mask size of 2X2, 3X3 and 4X4 respectively. It is observed that when the mask size is moderate then the output image will appear to be clear, and use of a low mask or higher size results in the blurred image as in Figure 3.2(b) and (d).

Unsharp Masking (USM) filter, on the other hand is a sharpening filter which enhances the blurred input image. Figure 3.3 shows the results of USM filter for sharpening the input image. The results of applying USM filter for a sample image in Figure 3.3 (a) using mask size of 2X2, 3X3 and 5X5 are shown in Fig 3.3 (b), (c) and (d) respectively. If the mask size is 2X2 the image is more sharpened and as the mask size increases the image is less sharpened resulting in a good output image.



(a) Camera Captured Input Epigraphic image

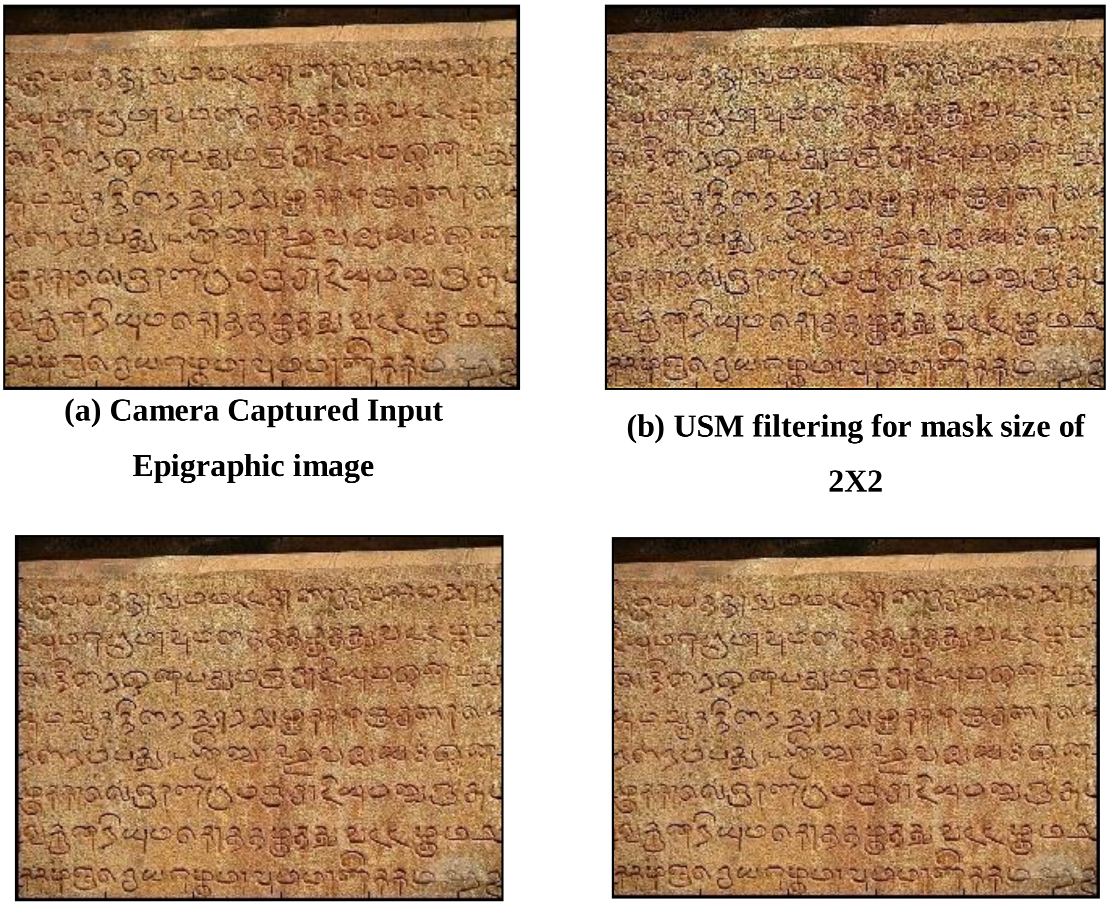
(b) Gaussian blurring for mask size of 2X2

(c) Gaussian blurring for mask size of 3X3

(d) Gaussian blurring for mask size of 4X4

Figure 3.2: Results of Gaussian blur Filtering on Camera Captured Epigraph

The Laplacian filter is another smoothing filter. Figure 3.4 shows the results of Laplacian filter for smoothing the input image. Figure 3.4 (b), and (c) shows results of applying a Laplacian filter on the input image shown in Figure 3.4 (a) using the diffusion value of -0.01 and +0.01 respectively. For D-value of -0.01, the input image is slightly sharpened and as the D-value increases the image smoothened. As a result for D-value of +0.01, the input image smoothened. As compared to other filters, Laplacian filter shows less difference in output compared to input.



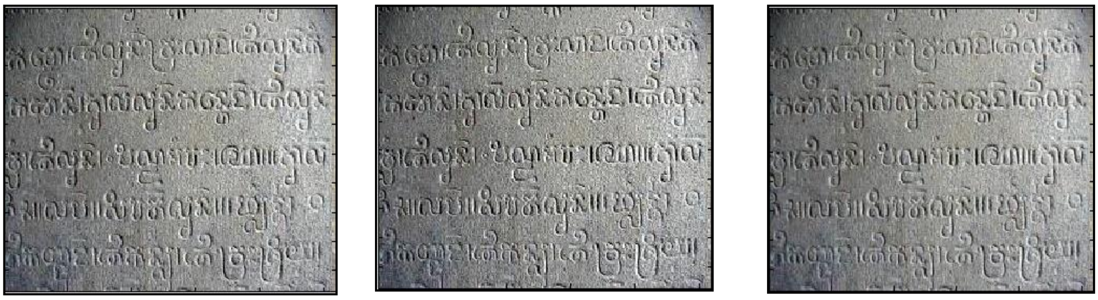
(a) Camera Captured Input Epigraphic image

(b) USM filtering for mask size of 2X2

(c) USM filtering for mask size of 3X3

(d) USM filtering for mask size of 5X5

Figure 3.3: Results of USM Filtering on Camera Captured Epigraph



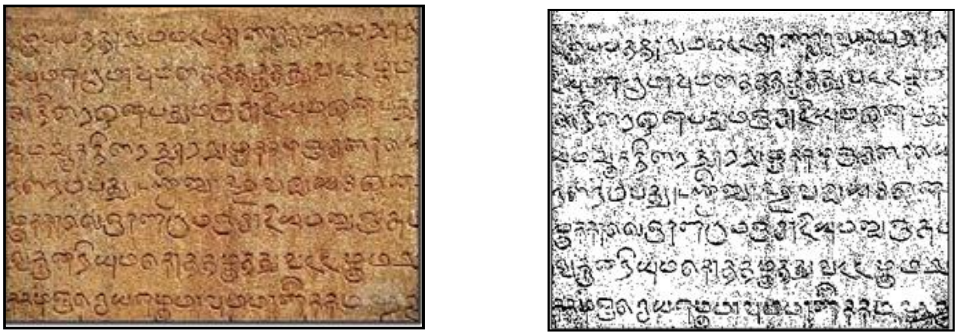
(a) **(b)** **(c)**

(a) Camera Captured Input Epigraphic image; (b)Laplacian filtering for D-value of -0.01, (c) Laplacian filtering for D-value of +0.01

Figure 3.4: Results of Laplacian Filtering on Camera Captured Epigraph

The smoothing filters used in this system will result in good accuracy if the edges of the input image are very thick, where as in the case of sharpening filter namely, USM filter better output is achieved for the degraded input image. The two smoothing filters used in the proposed system are a Gaussian blur and Laplacian filter, depending on the input image the required filter can be used with different mask sizes to get the filtered image. The enhancement is found to be appreciable for mask size of 3X3 for Gaussian blur, 5X5 for USM filter and in the case of Laplacian filter input image is sharpened for negative values of the diffusion constant, smoothened for positive values.

The enhanced image is binarized using Otsu's method and the result is shown in Figure 3.5.



(a) Input Epigraphic Image **(b) Binarization Results**

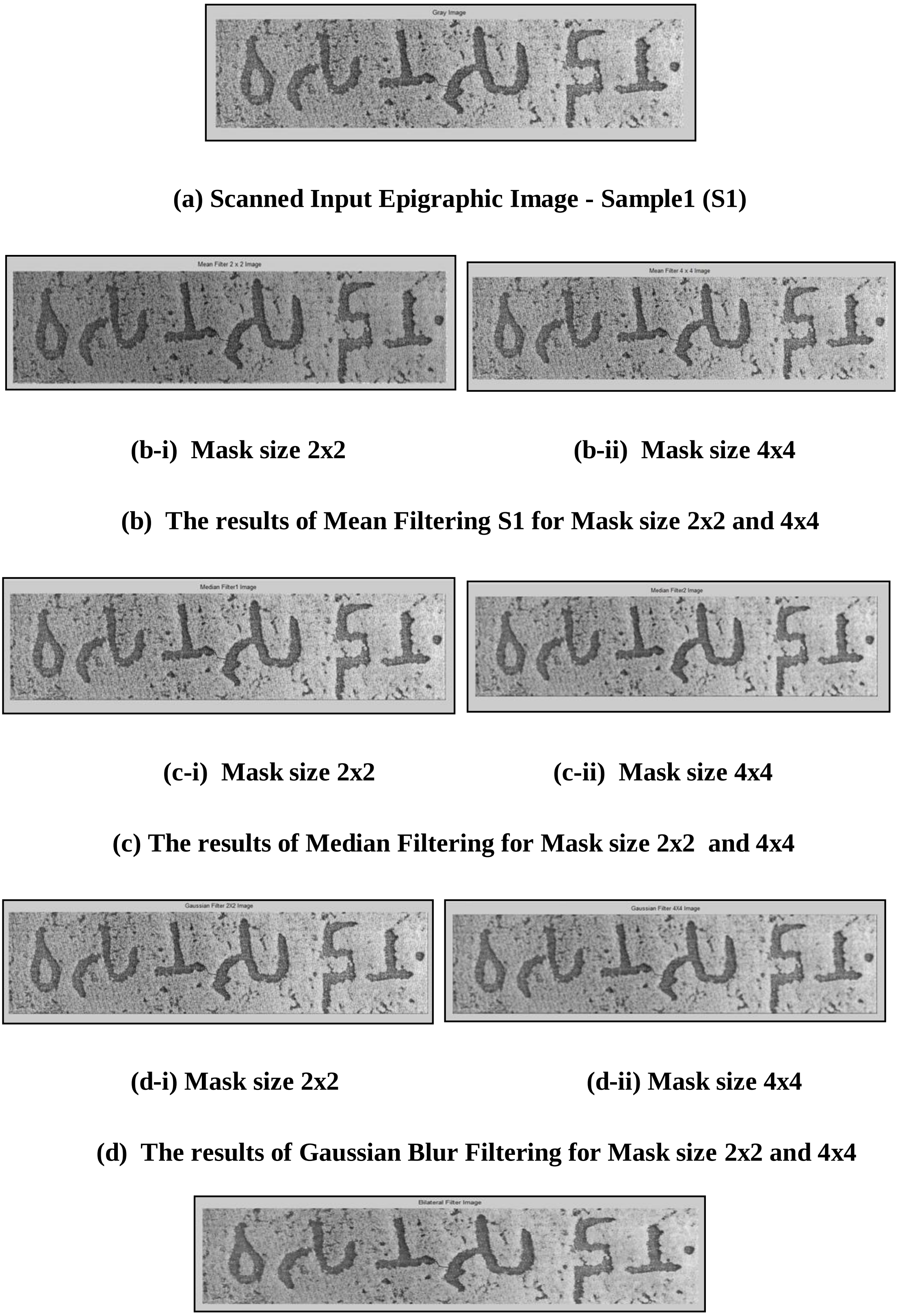
Figure 3.5: Result of Binarization using Otsu's method

3.6.3 Enhancement of Scanned Epigraphic Document Images

In this section, the experimental results and analysis of the image enhancement using spatial filtering techniques namely Mean, Median, Gaussian blur and Bilateral filter for different mask sizes are discussed with illustrations.

- **Experimental Results**

The spatial filtering techniques are explored on nearly 150 scanned ancient epigraphic images of varying image quality and degradations. Figure 3.6 and 3.7 shows the input ancient historical record with varying- writing material and degradation, which is analyzed for different spatial filtering techniques.



(a) Scanned Input Epigraphic Image - Sample1 (S1)

(b-i) Mask size 2x2 **(b-ii) Mask size 4x4**

(b) The results of Mean Filtering S1 for Mask size 2x2 and 4x4

(c-i) Mask size 2x2 **(c-ii) Mask size 4x4**

(c) The results of Median Filtering for Mask size 2x2 and 4x4

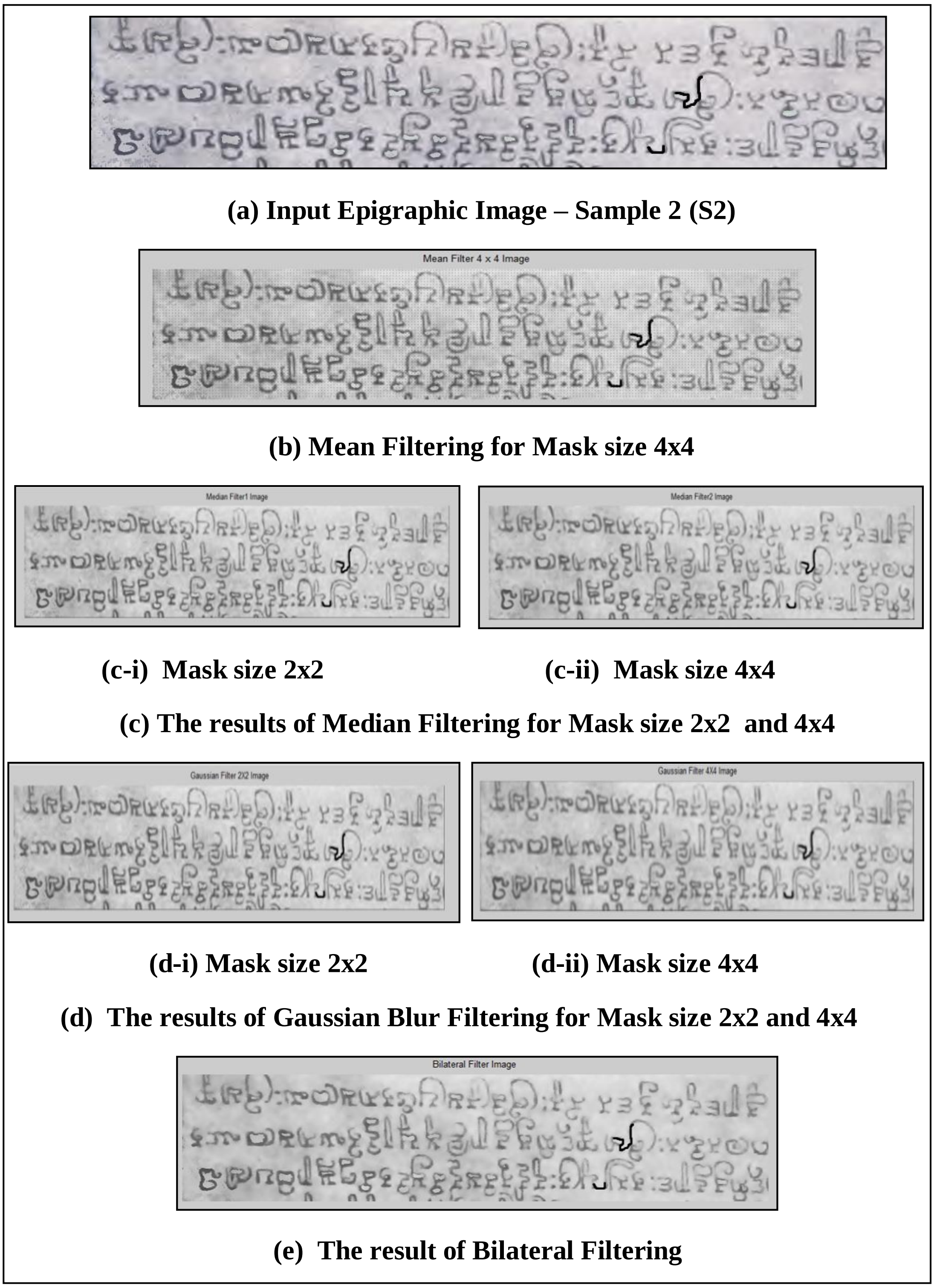
(d-i) Mask size 2x2 **(d-ii) Mask size 4x4**

(d) The results of Gaussian Blur Filtering for Mask size 2x2 and 4x4

(e) The result of Bilateral Filtering

Figure 3.6: Results of Spatial Filtering Scanned Epigraphic Image (S1)

Figure 3.7(a) shows the input epigraphic image Sample S2 subjected to various spatial filtering techniques in proposed model and results are depicted in Figures 3.7(b) to 3.7(e).



(a) Input Epigraphic Image – Sample 2 (S2)

(b) Mean Filtering for Mask size 4x4

(c-i) Mask size 2x2 **(c-ii) Mask size 4x4**

(c) The results of Median Filtering for Mask size 2x2 and 4x4

(d-i) Mask size 2x2 **(d-ii) Mask size 4x4**

(d) The results of Gaussian Blur Filtering for Mask size 2x2 and 4x4

(e) The result of Bilateral Filtering

Figure 3.7: Results of Spatial Filtering Scanned Epigraphic Image (S2)

- **Performance Analysis of Filtering Approaches**

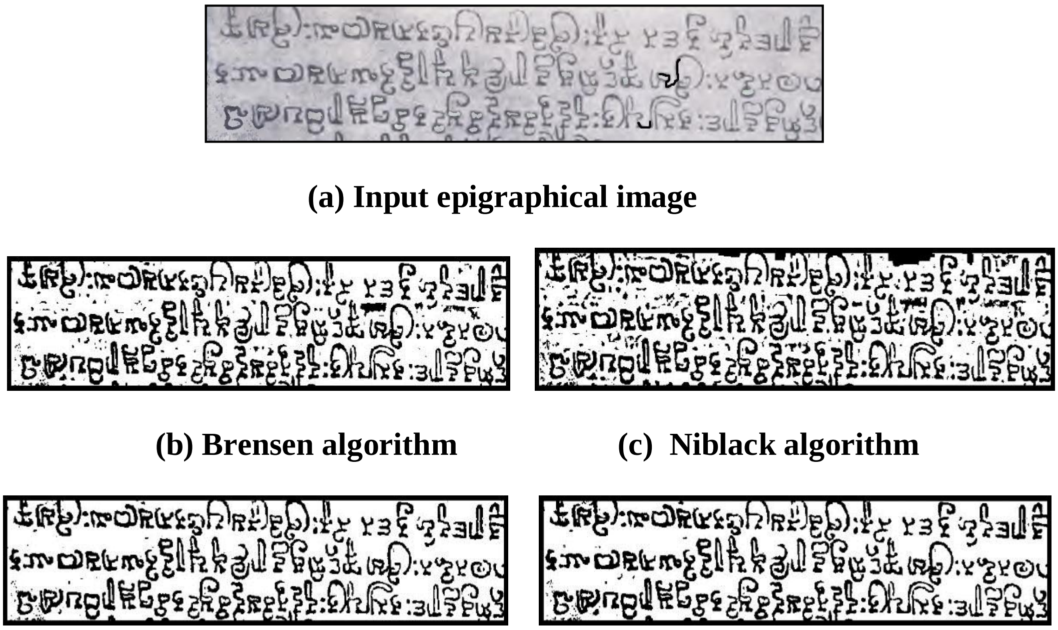
The performance of different spatial filtering techniques explored for enhancement of epigraphic images is elucidated in this section. The noise reduction and enhancement is tested on 150 scanned epigraphic samples of medium degradation, using four spatial filtering techniques mean, median, Gaussian blur and Bilateral of varying mask sizes.

The Mean filter with a mask size of 4x4 typically smoothens local variations in an image and noise is reduced as a result of blurring. Thus, it smears noise specks but also blurs the edges. The enhancement is found to be appreciable for the mask size 4x4 for Median filter when the mask size is high then the output image will appear clear with sharp edges. The median filter is an effective method that can suppress isolated noise without blurring sharp edges or preserves sharp edges. The mask size of 2x2 for Gaussian blur results in better blurred effect eliminating the background noise. Bilateral filter sharpens the edges. The smoothing Gaussian filter will result in good accuracy if the edge of the input image is very thick, whereas in the case of sharpening, Bilateral filter gives better output for the medium degraded images.

3.6.4 Binarization of Scanned Epigraphic Document Images

Various binarization techniques Brensen, Niblack, Sauvola and Savakis are tested on scanned epigraphic images and the results are compared. This is illustrated for the sample input epigraphic document shown in Figure 3.8(a). Figure 3.8(b)-3.8(e) shows the result of the binarization using Brensen, Niblack, Sauvola and Savakis algorithm respectively.

All the above algorithms for binarization are tested on 150 epigraphic images of a medium amount of degradation. The Sauvola algorithm performs comparatively better than Brensen, Niblack and Savakis for most of the images.



(a) Input epigraphical image

(b) Brensen algorithm

(c) Niblack algorithm

(d) Sauvola algorithm

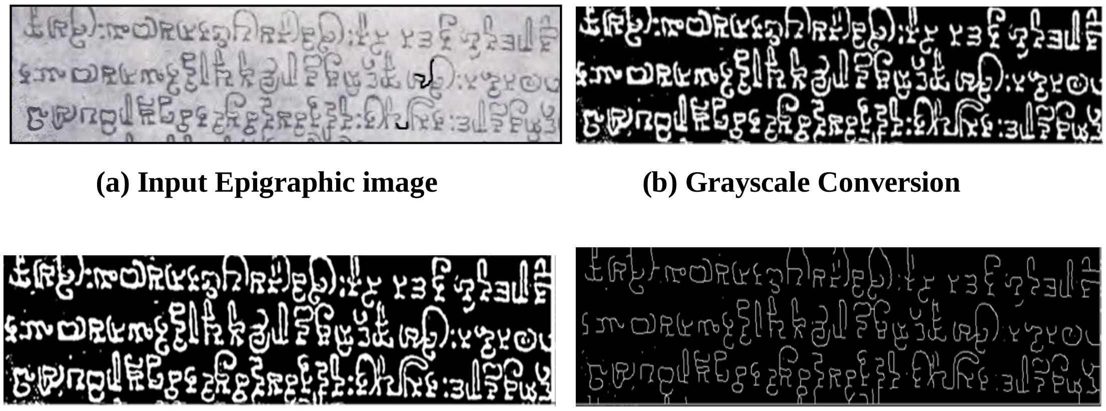
(e) Savakis algorithm

Figure 3.8: Results of Binarization using different approaches

3.6.5 Preprocessing of Scanned Epigraphic Document Images

This section illustrates the results of preprocessing a sample epigraphic document image depicted in Figure 3.9(a). Preprocessing involves the steps: gray-scale conversion, Contrast enhancement using Yu and Bajaj’s technique, Sharpening using Unmask filtering technique and Thinning using Guo-Hall method. The results of these steps are shown in Figure 3.9(b)-3.9(d) respectively.

Figure 3.9 (b) shows the result of the gray-scale conversion. The image is blurred by a small amount and hence is fed to the contrast enhancement and sharpening procedure. Figure 3.9(c) shows the result of the contrast enhancement and sharpening of the grayscale image. The output image highlights the edges of each character in the text. Figure 3.9(d) shows the thinned image in which the characters are thinned using the Guo Hall algorithm.



(a) Input Epigraphic image **(b) Grayscale Conversion**

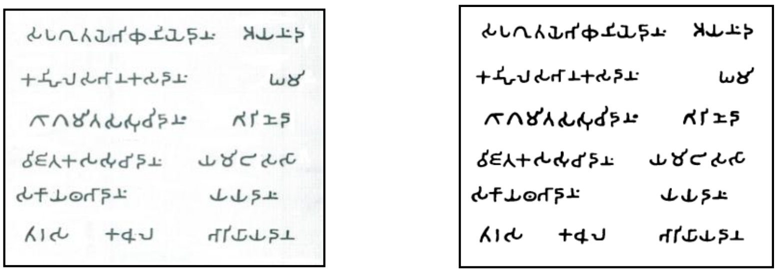
(c) Contrast Enhancement and Sharpening **(d) Thinning**

Figure 3.9: Preprocessing of Scanned Epigraphic Image (S2)

Thinning helps to improve the efficiency of processing and assist in next step of obtaining sampled characters to a certain extent by addressing the touching characters.

3.6.6 Preprocessing of Scanned Epigraphs of Good Quality

The preprocessing sequence that will address the scanned epigraphic images of good quality is presented here. The scanned input epigraphic image is shown in Figure 3.10(a) and the Figure 3.10(b) illustrates the final Pre-processed image which is the outcome of Spatial filtering namely Gaussian, Median and Bilateral followed by morphological operations erode operation and dilation operation.



(a) Input Epigraph **(b) Preprocessed Image**

Figure 3.10: Result of Preprocessing Scanned Epigraphic Image (S3)

The Preprocessing steps are tested on 50 scanned ancient epigraphic images with minimal noise and it is found that the combination of the three filters used namely Gaussian, Median and Bilateral, remove the noise from the image while preserving the edges. The morphological operations namely erode and dilate improves the clarity of edges in turn textual writing as a whole, thereby further improving the accuracy of segmentation and recognition.

3.6.7 Segmentation Approaches – Results and Performance Analysis

The segmentation techniques in the proposed work are tested on nearly 150 noise-free epigraphic document images. This section illustrates the experimental results and highlights the inferences drawn in testing these approaches for segmentation of epigraphic records.

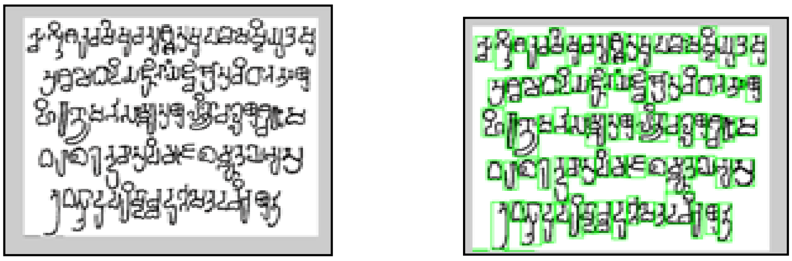
- **Evaluation Metric**

The metric used to evaluate the accuracy of Segmentation phase is the Segmentation Rate given by the Equation 3.1

$$\text{Segmentation Rate} = \frac{\text{Number of Correctly Segmented Characters}}{\text{Total Number of Characters in input document}} \qquad (3.1)$$

- **Connected-Component Labeling and Bounding box technique**

The segmentation algorithm Connected component and bounding box, is tested on the text of ancient periods. The results of character segmentation of sample ancient epigraphic image are demonstrated in Figure 3.11



(a) Input Epigraphic Image **(b) Segmentation Results**

Figure 3.11: Segmentation Results using Connected Component and Bounding Box Method - Sample (S4)

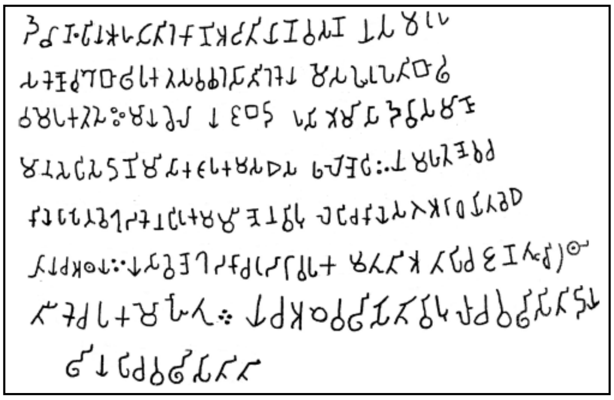
Performance Analysis

Segmentation using Connected Component and Bounding Box Method is tested on nearly 150 samples of reconstructed epigraphic images from different periods. The accuracy of the result is mainly dependent on appropriate connectivity of the pixels which forms the complete character in the document. It segments the base characters and compound characters correctly when the character sub-components have complete pixel connectivity. Few cases where in connectivity is not present, the compound character is segmented separately and output as disjoint characters. If the subcomponents that form a compound character are connected, then the character is segmented correctly. If there is disconnectivity of the character components, the output of segmentation results into two disjoint character components. If two characters are touching or overlapping then the characters are considered as a single letter and segmented incorrectly as a single character.

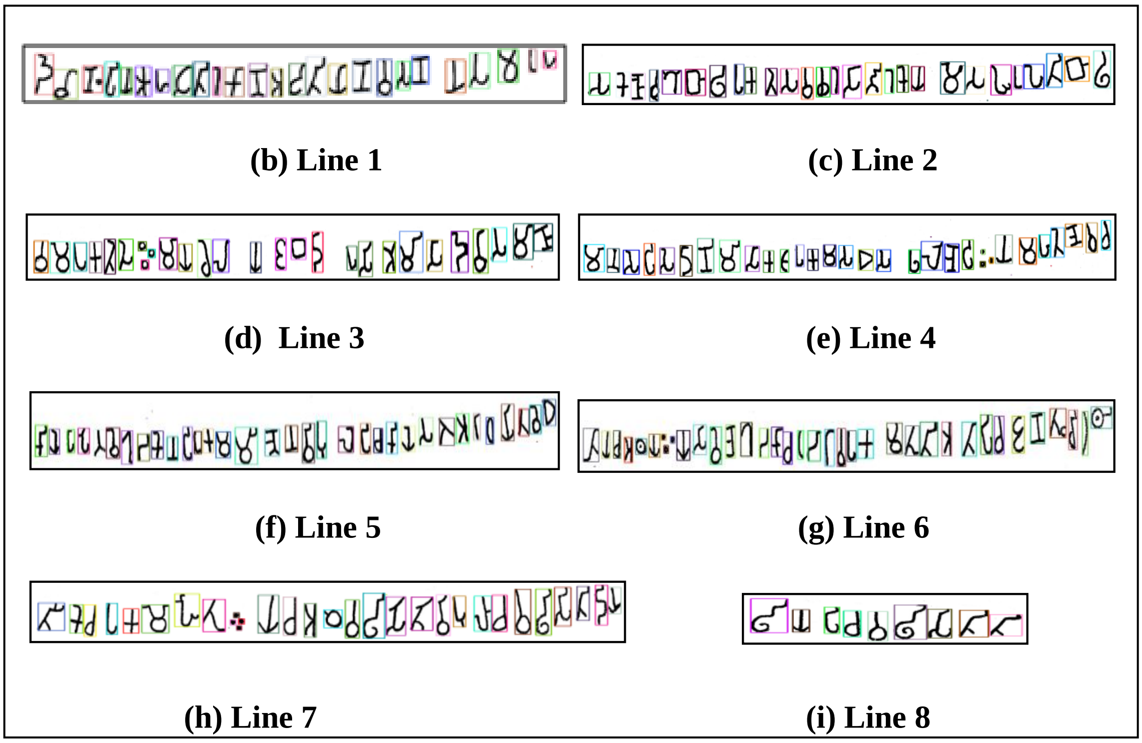
The accuracy rate of the segmentation around 83.5% is achieved for Brahmi script and 67.5% for other ancient Kannada scripts. The result for segmentation of Brahmi script is appreciable when compared to other ancient Kannada script from different periods because of the nature of the script. The complexity of writing style in Brahmi script is less compared to the other scripts from different times.

- **Contour based Convex Hull and Bounding Box technique**

Figure 3.12(a) is the reconstructed input epigraphic image. The line segmentation is performed to obtain individual lines of text and next is subjected to character segmentation to sample out the characters. The results of character segmentation after line segmentation are shown in Figure 3.12(b) – 3.12(i) using Contour Detection and Minimum Bounding Box approach. The characters are segmented irrespective of the skew present in the document.



(a) Reconstructed Input Epigraphical Document Image (S5)



(b) Line 1

(c) Line 2

(d) Line 3

(e) Line 4

(f) Line 5

(g) Line 6

(h) Line 7

(i) Line 8

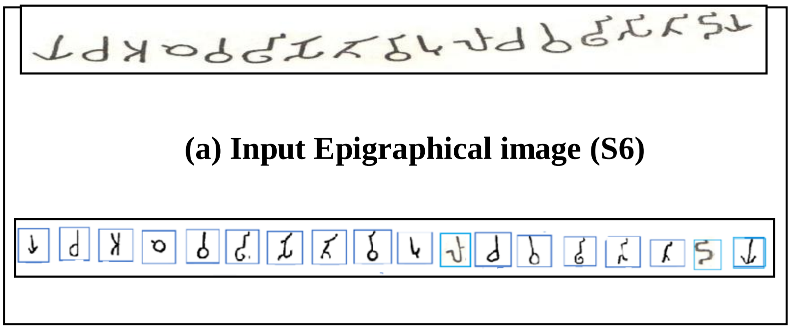
Figure 3.12: Results of Line and Character Segmentation using Contour Detection and Minimum Bounding Box Method

Performance Analysis

The Contour based Convex Hull and Bounding Box segmentation algorithm is tested on 150 epigraphical images of different periods each with varying number of characters. It works well for a noise free image both for base and compound characters irrespective of the size of the image and skew present. The algorithm is tested on pre-processed images of Brahmi script from Ashoka period. The spacing of text lines and the nature of characters reduces the complexity in the segmentation process of Ashokan Brahmi script yielding encouraging results of an average segmentation rate 92%. However, for the segmentation of other ancient Kannada scripts, achieved a segmentation rate of 83%.

- **Nearest Neighbor Clustering method**

Figure 3.13 depicts the result of segmentation using Nearest Neighbor Clustering. The method segments the image into lines and characters.



(a) Input Epigraphical image (S6)

(b) Segmentation of Characters

Figure 3.13: Segmentation Results of Nearest Neighbor Clustering Method

Performance Analysis

The nearest neighbor clustering based method segments the line and character and works even on the skewed document. The method segments the epigraph into lines and characters. The technique works fine for the skewed document. The result of the segmentation is dependent on the cuts and bruises in the input script image. A Segmentation rate around 87% is achieved for Brahmi script and 83.5% for other ancient Kannada scripts. The segmentation results for Brahmi script are appreciable when compared to other scripts because of the less complexity nature of characters.

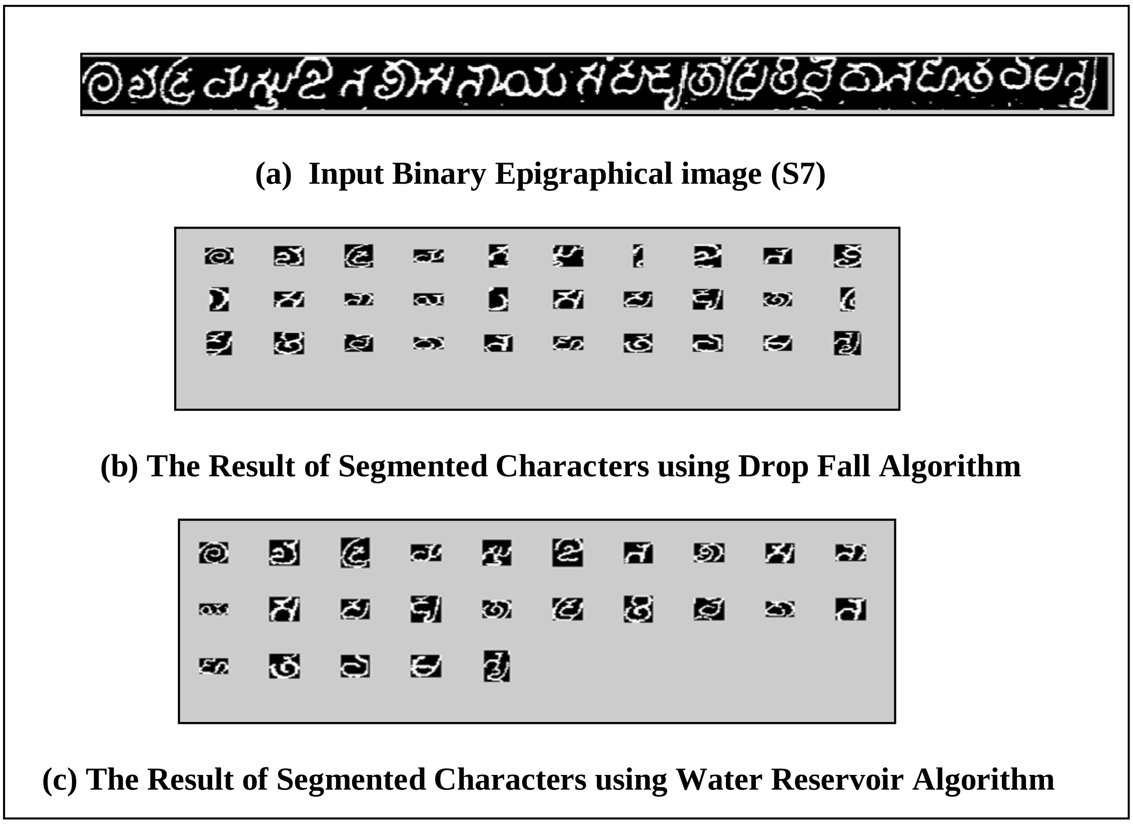
➢ Drop Fall and Water Reservoir Segmentation Techniques

Segmentation is also carried out using Drop Fall and Water Reservoir algorithms especially to address touching characters. A preprocessed and binarized epigraphical image shown in Figure 3.14(a) is considered as input to segmentation phase.

Figure 3.14(b) and Figure 3.14(c) represents the result of Segmentation of Characters using Drop Fall algorithm and Water Reservoir algorithm respectively.

Performance Analysis

Segmentation is carried out using Drop Fall and Water Reservoir algorithms to address touching characters. The techniques segment the base and compound characters correctly when the connectivity is present. Few cases, where in the subcomponents of a compound character are disjoint, the compound character is segmented separately into two components. These techniques work well preferably for segmentation of touching characters. A segmentation rate of 85.6% for Drop Fall algorithm and 86.7% for Water Reservoir algorithm is achieved when tested on 150 preprocessed epigraphic documents.



(a) Input Binary Epigraphical image (S7)

(b) The Result of Segmented Characters using Drop Fall Algorithm

(c) The Result of Segmented Characters using Water Reservoir Algorithm

Figure 3.14 : Result of Segmented Characters using Drop Fall and Water Reservoir Algorithms

3.7 Summary

In this chapter importance of preprocessing and segmentation of epigraphical documents have been discussed and techniques are proposed to address the same.

Preprocessing includes the enhancement of epigraphical records, noise elimination and binarization. The experimental results of preprocessing which involves enhancement and noise removal stages are tested on 250 camera-grabbed and scanned epigraphical images. Ancient epigraphical documents are enhanced by using suitable Spatial filtering techniques. Mean, Median, Gaussian Blur, Bilateral, Laplace filter, Unsharp Masking (USM) filters are explored with different filter sizes and filter parameters. Thus, the enhancement of historical records - transforms the degraded input document into a better perceivable image. The system performs well for preprocessing of ancient documents and provides flexibility to the user in controlling the process of image enhancement to obtain desired output. Next different binarization techniques Otsu's, Brensen, Niblack, Sauvola and Savakis are explored and observed that Sauvola performs better than others.

The next phase, Segmentation of epigraphical documents is carried out based on Connected Component Labeling and Bounding Box technique, Contour based Convex Hull and Minimum Bounding Box approach, Nearest Neighbor Clustering, Drop Fall and Water Reservoir techniques. Connected Component and Bounding Box method segments characters of reconstructed epigraphical images correctly, when there exists complete connectivity in the character. The method fails for the compound character with disconnected sub-components. Contour based Convex Hull and Minimum Bounding Box performs well both for base and compound characters yielding better segmentation results irrespective of skew in the image. Nearest Neighbor Clustering is used to perform segmentation of lines and characters successfully. The technique works fine for the skewed document. Segmentation of documents with touching characters is performed using Drop Fall and Water Reservoir algorithms. The output of segmentation phase is the sampled characters which are used in the next stages for Classification and Recognition of epigraphic characters.

Chapter 4

CLASSIFICATION OF ANCIENT KANNADA EPIGRAPHS INTO DIFFERENT PERIODS

4.1 Premise

Ancient handwritten script identification is a complex task due to the complexities in pre-processing, feature extraction and classification, the sensitivity of the technique to the variation of handwritten text in the document (font style, font size and document skew) and performance of the technique. In the previous chapter, various available approaches for preprocessing and segmentation of epigraphic records are explored. It is observed that the characters of a script have evolved overtime and different periods have different character set [1]. Hence it is essential to determine the period of epigraph before deciphering the script. This chapter presents SVM model and the design of Random Forest (RF) Classifier for predicting the period of epigraphic records.

The organization of the chapter is as follows:- The need towards the classification of epigraphs into different periods is highlighted in Section 4.2. The proposed model is presented in Section 4.3. The related theory and background of the approaches used in the proposed work are covered in Section 4.4. Section 4.5 covers in detail the design and the algorithms for dating epigraphical records. Experimental results and performance analysis are discussed in Section 4.6. Finally, Section 4.7 summarizes the chapter.

4.2 Importance of Period Identification

The ancient scripts have evolved overtime and transformed to present form. Hence, scripts across different periods have different character set. During the evolution of the Kannada script from 3rd century B.C to 18th century A.D as depicted in Figure 4.1, there is a noticeable change in the shape of the characters [1]. This poses a challenge in reading these ancient scripts. The prediction of the period of the script is essential to know the character set to be used for deciphering. Hence, the period of the script to which it belongs is to be estimated before the actual character recognition is taken up.

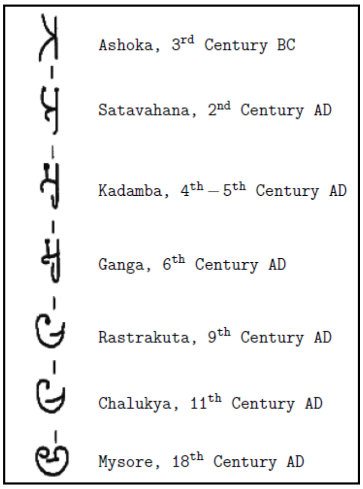

Ashoka, 3rd Century BC
Satavahana, 2nd Century AD
Kadamba, 4th – 5th Century AD
Ganga, 6th Century AD
Rastrakuta, 9th Century AD
Chalukya, 11th Century AD
Mysore, 18th Century AD

Figure 4.1: The Evolution of Character ‘a’ from 3rd century BC to 18th century AD

(Courtesy: “Kannada Lipiya Ugama Mattu Vikasa,” by A. V. Narasimha Murthy)

The dating of inscriptions, reading and translation of the epigraphic documents is a difficult and time consuming process if done manually by epigraphists. Hence, there is a need for an automated system to perform the tasks of epigraphers. It is necessary to first determine the period of the epigraph, since the process of building a system for automatic reading of ancient documents is made easier, by reducing the dictionary of characters to be used for the recognition. The system thus developed assist archaeologists in dating scripts inscribed on objects. It also facilitates many important applications such as automatic archiving of multilingual documents, searching online archives of document images of particular era and for the selection of the script specific OCR in an environment where scripts of various periods are given as input.

4.3 Proposed Model for Period Identification

Period identification of ancient scripts is important, which enables to know the character bank to be employed for automatic reading of age-old documents. An ancient Kannada epigraph is input to the system and output is the predicted era of the script.

The generic Period Identification model is shown in Figure 4.2 and involves the following components:

- **Preprocessing:** This sub-system pre-processes and enhances the input document image before it is passed to further stages.

- **Segmentation:** Sampled characters are extracted from the document images which are normalized to a particular size. The characters thus obtained are fed into the Feature Extraction phase.

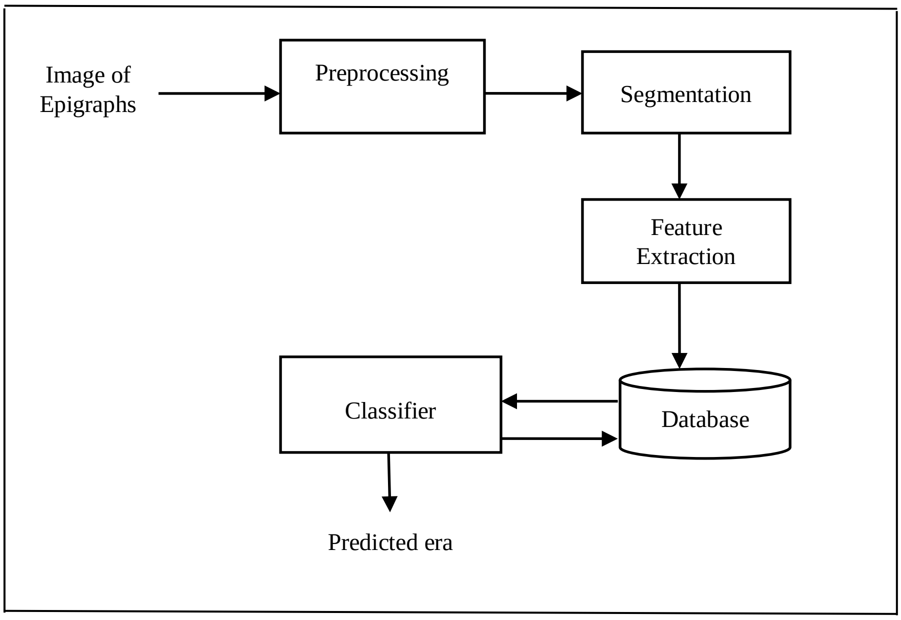


Figure 4.2: Proposed Model for Period Identification

- **Feature Extraction:** The distinguishing features are extracted from the sampled characters and saved in a file.

- **Database:** The database here represents the file used to save the extracted features. The feature vectors are used for training the classifier and later during testing for the prediction of the era.

- **Classifier:** The Classifier is trained using the features stored in the file. It is also possible to save the trained Classifier for later use. The trained Classifier is used further to predict the era.

The proposed work uses two models for predicting era of ancient Kannada epigraphs. The first model is SVM Classifier with Zone-based features and the other is Random Forest Classifier (RF) with Central and Zernike moments.

4.3.1 SVM Classifier with Zone-based features

The prediction of the era is done by examining few characters in ancient Kannada script of various periods referred to as unique characters. Initially, a data corpus of unique characters (which are characters that are distinct and non-repetitive from one era to the other) is created. Zone-based features are extracted from these unique characters and used as the feature vectors during training the classifier. Later during testing, the system considers a preprocessed epigraphic script as input, segments it into individual characters and then extracts Zone-based features for the segmented characters and classifies them based on the period. The count of the character class labels belonging to different periods is determined. Finally, the prediction of the period of the script is made on the basis of the majority count.

4.3.2 RF Classifier with Normalized Central and Zernike Moments

Ancient Kannada epigraphs of different periods are considered as input. The image is preprocessed and segmented into characters. Normalized Central Moments and Zernike Moments are extracted from the segmented characters and used as the feature vectors for classification. There are 9 features as Central Moments and 14 features as the Zernike Moments. Random Forest is used as the classifier, which is an ensemble of classification trees, and each tree votes for a class and the output class is the majority of the votes. The class labels used is the eras of the document images. After all the characters in the image are classified, the majority of the classes output is shown as the era of the epigraph image.

4.4 The Related Theory and Background

This section discusses the theory and background related to the present work on Prediction of the Period of input epigraphical script.

4.4.1 SVM Model with Zone-based features

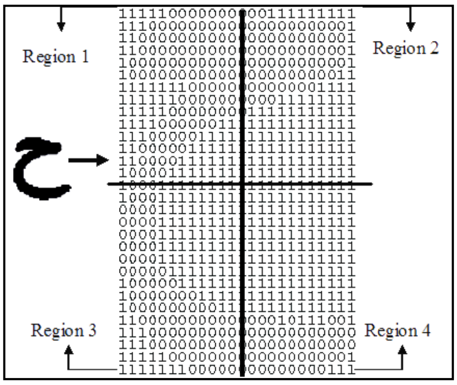
- **Data Corpus Creation:** The unique set of characters, which distinguishes it from other eras and which is found regularly in the scripture are chosen for different periods. Thus the instances of these unique characters from different periods form the data corpus. The use of unique characters in the dataset reduces the search space, thereby reducing the complexity in identifying the script.
- **FeatureExtraction:** Zone-based features are extracted for the classification process.

There are totally14 features extracted from the character of which 4 of them are for the whole image as listed below:

Height / Width, Number of black pixels / number of white pixels image, Number of horizontal transitions, Number of vertical transitions.

In addition, the image is divided into four regions as shown in Figure 4.3 and the Number of Black Pixels /Number of White Pixels across horizontal regions, vertical regions and diagonal regions is extracted.

Also, pixel density, mean, variance, standard deviation, the transition of black to white pixels is computed for each character from the script and forms the database. The features extracted are independent of the font size of the characters. These features are also easy to extract and distinct in nature thereby reducing the complexity of the system. The system is trained with the features of the characters present in the database and later used for classification.



```
111110000000 001111111111
111000000000 000000000001
110000000000 000000000001
110000000000 000000000001
100000000000 000000000001
100000000000 000000000011
111111100000 000000001111
111111000000 000111111111
111110000000 111111111111
111100000011 111111111111
111000001111 111111111111
110000011111 111111111111
110000111111 111111111111
100001111111 111111111111
100011111111 111111111111
100011111111 111111111111
000011111111 111111111111
000011111111 111111111111
000011111111 111111111111
000011111111 111111111111
000011111111 111111111111
000001111111 111111111111
100000111111 111111111111
100000001111 111111111111
100000000011 111111111111
110000000000 000010111001
111000000000 000000000000
111000000000 000000000000
111110000000 000000000001
111111100000 000000000111
```

Figure 4.3: Zone-based Feature Extraction by Dividing the Image into 4 Regions

- **Classification:** For the classification of the script to their respective period, a Support Vector Machine (SVM) Classifier using linear kernel function is used.

SVM is used as a multiclass-classifier system which is trained on the distinct / unique characters in the Kannada script of various periods. Each vector is given to SVM which classifies it into a class in the vector space.

The features are extracted for all the distinct characters of different periods. The unique characters identified a prior, for various periods are searched in the given input ancient script. The number of matches of these unique characters of the different era,

with the characters of input document image is accounted. The period of input epigraphical document image of ancient Kannada script is the one with a maximum number of matches of characters of input ancient script with unique characters.

Thus, the input document image is classified to the period of any of these dynasties: Ashoka dynasty, Satavahana dynasty, Kadamba dynasty, Badami Chalukya dynasty, Rashtrakuta dynasty, Kalyan Chalukya dynasty, Hoyasala dynasty, Vijayanagar dynasty or Mysore Wodeyar dynasty.

4.4.2 RF Model with Normalized Central and Zernike Moments

This section discusses the related theory with a mathematical background of statistical features such the Central Moments and Zernike moments [78], and the novel approach of designing RF classifier.

- **Central Moments**

For a continuous function *f(x, y)*, the Geometric Moment of order *(p+q)* is given in Equation 4.1:

$$M_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)dxdy \tag{4.1}$$

for *p*, *q* = 0, 1, 2… Adapting this to a grayscale image with pixel intensities *I(x, y)*, raw image Moments M_{ij} are calculated using Equation 4.2

$$M_{ij} = \sum_x \sum_y x^i y^j I(x,y) \tag{4.2}$$

From Equation (4.2), area is defined to be M_{00} and the centroid is given by Equation 4.3:

$$(\bar{x}, \bar{y}) = \left(\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}}\right) \tag{4.3}$$

Central moments are defined as:

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \bar{x})^{p}(y - \bar{y})^{q} f(x, y)\, dx\, dy \tag{4.4}$$

where $(\bar{x}, \overline{y})$ are the components of the centroid.

If *f (x, y)* is a digital image, then Equation 4.4 is written as given in Equation 4.5:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^{p}(y - \bar{y})^{q} f(x, y) \tag{4.5}$$

The Central Moments of the order up to 3 are given from Equation (4.6) to (4.15).

$$\mu_{00} = M_{00} \tag{4.6}$$

$$\mu_{01} = 0 \tag{4.7}$$

$$\mu_{10} = 0 \tag{4.8}$$

$$\mu_{11} = M_{11} - \bar{x}M_{01} = M_{11} - \bar{y}M_{10} \tag{4.9}$$

$$\mu_{20} = M_{20} - \bar{x}M_{10} \tag{4.10}$$

$$\mu_{02} = M_{02} - \bar{y}M_{01} \tag{4.11}$$

$$\mu_{21} = M_{21} - 2\bar{x}M_{11} - \bar{y}M_{20} + 2\bar{x}^{2}M_{01} \tag{4.12}$$

$$\mu_{12} = M_{12} - 2\bar{y}M_{11} - \bar{x}M_{02} + 2\bar{y}^{2}M_{10} \tag{4.13}$$

$$\mu_{30} = M_{30} - 3\bar{x}M_{20} + 2\bar{x}^{2}M_{10} \tag{4.14}$$

$$\mu_{03} = M_{03} - 3\bar{y}M_{02} + 2\bar{y}^{2}M_{01} \tag{4.15}$$

The Normalized Central Moments are computed by dividing the Central Moments by M^2_{00} in order to keep it invariant to scale.

- **Zernike Moments**

In the Polar co-ordinate system, where (ρ, θ) are the co-ordinates, the Zernike radial polynomials $\{R_{nm}(\rho)\}$ are given by Equation 4.16:

$$R_{nm}(\rho) \;=\; \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s \frac{(n-s)!\rho^{n-2s}}{s!\left(\frac{n-|m|}{2}-s\right)!\left(\frac{n+|m|}{2}-s\right)!} \tag{4.16}$$

where n is a non-negative integer, and m is a non-zero integer subject to the following constraints: n – │m│ is even and │m│≤ n.

A set of complex polynomials is introduced which form a complete orthogonal set over the interior of the unit circle, i.e. $\rho^2 = x^2+y^2 \leq 1$. They have a form given by Equation 4.17:

$$V_{nm}(x,y) = V_{nm}(\rho,\theta) = R_{nm}(\rho)\exp(j\;m\;\theta) \tag{4.17}$$

where

n – Positive integer or zero.

m – Positive and negative integers subject to constraints

n –│m│ (even), │m│≤ n

ρ – Length of the vector from the origin to (x, y) pixel, given by

$$\rho \;=\; \sqrt{x^2 \,+\, y^2}$$

q – Angle between vector ρ and x-axis in a counterclockwise direction, i.e.

$$\theta \;=\; \tan^{-1}\frac{y}{x}$$

These polynomials are orthogonal and Equation 4.18 holds good.

$$\int\int_{x^2+y^2\leq 1} r[V_{nm}(x,y)]^* \,\times\, V_{pq}(x,y)\,dx\,dy \;=\; \frac{\pi}{n+1}\delta_{np}\delta_{mq} \tag{4.18}$$

with

$$\delta_{ab} = \begin{cases} 1 & a = b \\ 0 & otherwise \end{cases}$$

Zernike Moments are defined as the projection of the image function onto these orthogonal basis functions. The Zernike Moment of order *n* with repetition *m* for a continuous image function *f (x, y)* that ceases to exist outside the unit circle is given by Equation 4.19:

$$A_{nm}(\rho, \theta) = \frac{n+1}{\pi} \int \int_{x^2+y^2 \leq 1} f(x, y)[V_{nm}(\rho, \theta)]^* \, dx \, dy \qquad (4.19)$$

and for the digital images is given by Equation 4.20:

$$A_{nm}(x, y) = \frac{n+1}{\pi} \sum_x \sum_y f(x, y)[V_{nm}(\rho, \theta)]^*, \quad x^2 + y^2 \leq 1 \qquad (4.20)$$

When computing the Zernike Moments of a given image, the center of the image is taken as the origin and the pixel coordinates are mapped to the range of unit circle, i.e., $x^2+y^2 \leq 1$. Those pixels falling outside the unit circle are not used in the computation.

The Zernike Moments computed are not invariant to scale and translation. Hence, in order to make it translation invariant, the formula is changed to:- $x_d = x - \overline{x}$, $y_d = y - \bar{y}$ and the values of the Translation Normalized Zernike Moments are given by $A_{nm}(x_d , y_d)$. For Scale Normalization, those values are divided by M_{00}. The Normalized Zernike Moments of order d = 6 are calculated, which means that the Zernike Moments calculated are: A_{20}, A_{22}, A_{31}, A_{33}, A_{40}, A_{42}, A_{44}, A_{51}, A_{53}, A_{55}, A_{60}, A_{62}, A_{64}, A_{66}.

➢ Classification Trees

Tree-based methods partition the feature space into smaller sets and then fit a simple tree to each one. They are conceptually simple yet powerful techniques [74, 75]. If the target is a classification outcome taking values 1, 2, … K, the tree algorithm needs to know the criteria for splitting the nodes.

In a node *m*, representing a region R_m with N_m observations, Equation 4.21 represents the proportion of class *k* observations in node *m*.

$$\hat{P}_{mk} = \frac{1}{N_m}\sum_{i \in R_m} I(y_i = k) \tag{4.21}$$

Classification is made to the observations in node *m* to class *k(m)* = arg$_k$ max $\hat{P}_{mk}$, the majority class in node *m*. The function *I* returns 1 if the condition is true. Different measures of node impurity are given by Equations 4.22 – 4.24:

Misclassification Rate:

$$\frac{1}{N_m}\sum_{i \in R_m} I\big(y_i \neq k(m)\big) = 1 - \hat{P}_{mk(m)} \tag{4.22}$$

Gini Index:

$$\sum_{k \neq k^j} \hat{P}_{mk}\hat{P}_{mk^j} = \sum_{k=1}^{K} \hat{P}_{mk}\big(1 - \hat{P}_{mk}\big) \tag{4.23}$$

Cross-entropy:

$$-\sum_{k=1}^{K} \hat{P}_{mk} \log \hat{P}_{mk} \tag{4.24}$$

E.g. When there are two classes, if p is the proportion in the second class, these three measures are 1- max(p, 1-p); 2 p(1-p) and -p log p - (1- p) log(1- p), respectively.

Measures of node impurity are shown in Figure 4.4. All three are similar, but cross-entropy and the Gini index are differentiable, and hence more amenable to numerical optimization. In addition, cross-entropy and the Gini index are more sensitive to changes in the node probabilities than the misclassification rate. For example, in a two-class problem with 40 observations in each class (denote this by (40, 40)), suppose one split created nodes (30, 10) and (10, 30), while the other created nodes (20, 40) and (20, 0). Both split produce a misclassification rate of 0.25, but the latter split produces a pure node and is likely preferable. Both Gini index and cross-entropy are smaller for the second split. Hence, either Gini index or cross-entropy must be used when growing the tree.

- **Random Forest (RF) Classifier**

Random Forest (RF) is a kind of ensemble classifier, wherein the feature vector is given to several classifiers, and the class is determined by voting from the individual classifiers. In RF, the classifiers are the decision trees and it constructs a series of Classification Trees which will be used to classify a new example. For reducing the variance of an estimated prediction function of a Forest, a technique known as bagging or bootstrap aggregation is used. The idea used to create a classifier model is constructing multiple decision trees, each of which uses a subset of attributes randomly selected from the whole original set of attributes [76, 77].

Each tree is grown as follows:

1 If the number of cases in the training set is N, sample N cases at random with replacement from the original data (subset) will be the training set for growing the tree.

2 If there are M input features, a number $m <<M$ is specified such that at each node, m features are selected randomly out of the M and the best split on these m features is used to split the node. The value of m is held constant during the Forest growing.

3 Each tree is grown to the largest extent possible with no pruning.

The Forest error rate depends on two parameters listed below:

- The correlation in the Forest between any two trees. Decreasing the correlation decreases the Forest error rate.

- The strength of each tree in the Forest. A tree is a strong classifier if it has a low error rate. Decreasing the strength of the individual trees increases the Forest error rate.

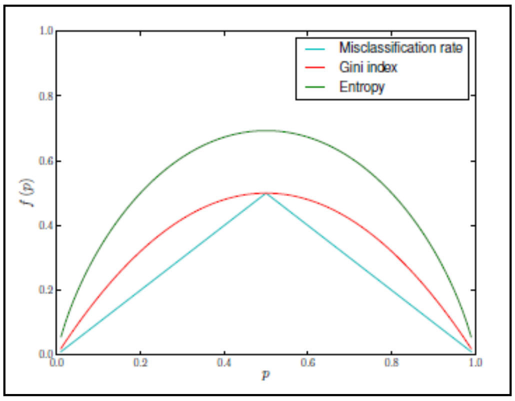


Figure 4.4: Comparison between Different Measures of Node Impurity for Two Classes

4.5 Detailed Design and Algorithms for Dating Epigraphs

This section presents in detail the two models SVM and RF for classification of ancient epigraphic records into different periods.

4.5.1 SVM Model with Zone-based features

Algorithm : **ERA_PREDICTION_SVM** (Epigraph_Image)

Input: Reconstructed ancient Kannada epigraph

Output: Classification of script into one of the following periods:

Ashoka, Satavahana, Kadamba, Badami Chalukya, Rashtrakuta, Kalyan Chalukya, Hoysala , Vijayanagar , Mysore Wodeyar dynasty.

Method:

Step 1: Identify unique characters (which are characters that are distinct and non-repetitive from one era to the other) of all the periods.

Step 2: [Create Data Corpus]: Create dataset containing unique characters from different eras. Instances of the unique characters from every era are created and stored in the database.

Step 3: [Feature Extraction]: Extract the features (pixel density, mean, variance, standard deviation, fraction of the black to white pixels) of unique characters in the dataset and store the values in the database, so as to use in the training phase.

Step 4: [Training]: Train the SVM model/classifier with the different instances of unique characters from each era.

Step 5: [Testing]

a. Extract the same set of features of all the characters from the test image and classify them into appropriate era using SVM model.

b. When the input character matches the unique character of a particular era, update the counter of the respective period appropriately.

Step 6: [Period Prediction]

a. Determine the count of the character class labels belonging to different periods.

b. The majority of the classified character class labels is the Period of the input epigraphic script.

4.5.2 RF Model with Central and Zernike features

***Algorithm*: PREDICT_ERA_RF (Epigraph_Image)**

Input: Reconstructed ancient Kannada epigraph

Output: Classification of script into one of the following periods:

Ashoka, Satavahana, Kadamba, Chalukya, Rastrakuta and Hoysala

Method:

Step 1: [Preprocessing]: Perform enhancement and noise removal.

Step 2: [Segmentation]: Perform character segmentation.

Step 3: [Feature Extraction]: Compute Normalized Central Moments and Normalized Zernike Moments, and write the feature vectors to a text file. There are 9 features as Central Moments and 14 features as the Zernike Moments.

Step 4: [Random Forest Classification]

a. **[Load Text]:** Get the feature vectors from the text file and save it in two arrays, one consisting of the classes and the other consisting of feature vectors of the corresponding classes.

b. **[Fit Forest]:** Train the trees in the RF which can be used to classify the ancient Kannada characters.

c. **[Fit Tree]:** A random subset of the training data from the step Fit Forest is taken as input and a single Classification Tree for the given subset of data is made.

d. **[Get Gini Impurity]:** Determine the impurity index of a subset of classes and corresponding data for the node so that it can find the best split and the best threshold value for that feature.

e. **[Predict]:** Considering the data consisting of feature vectors, predict the class of the given test characters using the trained RF Classifier.

Thus, the RF model has four functions to perform when building the Random Forest. The Data Subset Selection selects about 65% of the feature vectors with repetition, then m random features are selected to be used in the classification tree being built, after that, a specified number of random thresholds are chosen which are used for making a decision at the node. Using these functions, a specified number of classification trees are built to make a forest. These grown classification trees are then used for predicting the eras of epigraph images.

- **Algorithmic design of RF Model**

Random Forest Classification: This model classifies the ancient Kannada script into any of these periods: Ashoka, Satavahana, Kadamba, Chalukya, Rastrakuta and Hoysala.

Step 1: [Load Text]

Gets the feature vectors from the text file and saves it in two arrays, one consisting of the classes and the other consisting of feature vectors of the corresponding classes.

Step 2: [Fit Forest]: Train the trees in the forest

Algorithm**: FOREST_**FIT (y_train, x_train)

Input: Arrays extracted from the Step- Load Text

Output: A trained RF which can be used to classify the ancient Kannada characters

Method:

```
        models ← an empty list
        ntrees ← 10
        for i in 1 to n trees do
            label_subset ← 65% of y_train
            data_subset ← data corresponding to label_subset
            append TREE_FIT (label_subset, data_subset) to models
        end for
 end method
```

Step 3: [Fit Tree]: Make a Classification Tree for the given input data.

Algorithm**:** TREE_ FIT (labels, data)

Input: A random subset of the training data from the sub-module Fit Forest

Output: A single Classification Tree for the given subset of data.

Method:

```
    classes ← unique values in the labels
    nclasses ← the number of classes
    nrows, nfeatures  ← the number of rows and columns in data
```

```
        m ← [ lg(nfeatures)+ .5⌋
        root ← SPLIT (labels, data, m, 1)
    end method
```

Step 3a: [SPLIT]

```
    Algorithm: SPLIT (labels, data, m, height)
    if nsamples ≤  min_leaf_size or height >max_height then
        nleaves  ← nleaves + 1
        return majority of labels
    else
        rand_feature_indices ←  m features randomly selected from data
        best_ split_score ← ∞
        for each feature_index in rand_feature_indices do
            feature_vec  ←  array of features from feature_index column
            thresholds  ← the midpoint values from feature_vec
            Find threshold and combined_score by using BEST_GINI_SPLIT
            if combined score < best_split_score then
                best_split_score  ←  combined_score
                best_feature_index  ← feature_index
                best_threshold  ←  threshold
            end if
        end for
        left_data ← data < best_threshold in best_feature_index
        left_labels  ←  the corresponding classes
        right_data   ← data ≥ best_threshold in best_feature_index
        right_labels  ←  the corresponding classes
        left_node  ← SPLIT(left_labels, left_data, m, height+1)
        right_node ← SPLIT(right_labels, right_data, m, height+1)
        node ← best_feature_index, best_threshold, left_node, right_node
        return node
    end if
    end method
```

Step 4: [Get Gini Impurity]: Determines the impurity index for a node with the data it has so that it can determine the best split feature and the best threshold value for that feature.

***Algorithm*: BEST_GINI_SPLIT(classes, feature_vec, thresholds, labels)**

```
    Input: A subset of classes and corresponding data for the node
    Output: The best feature and the best threshold
    Method:
    n ← number of labels
    for t in thresholds do
        left_labels ← labels having feature_vec < t
        right labels ← labels having feature_vec ≥ t
        left_score ← GINI (classes, left_labels)
        right_score ← GINI(classes, right_labels)
        nleft ← number of left_labels
        nright ← number of right_labels
        combined_score ← (nleft/n) left_score + (nright/n) right_score
        if combined_score < best_score then
            best_t ← t
            best_score ← combined_score
        end if
    end for
```

```
return best_t, best_score

Algorithm: GINI (classes, labels)

sum_squares ← 0
    n ← number of labels
    if n = 0 then
        return 0
    else
        for c in classes do
            count ← $\sum_{l \in labels}[l = c]$
            pp ← $count^2 / n^2$
            sum_squares ← sum_squares + pp
        end for
end if
return 1 – sum_squares
```

Step 5: [Predict]: Predicts the class for the given test characters using the trained Random Forest Classifier

```
Algorithm: PREDICT (x_test)

Input: Data consisting of feature vectors

Output: The predicted class

Method:
        Initialize vote_ y to an empty list
        for each x in x_test do
        Initialise pred_y to an empty list
        for each tree in models do
                Predict the class using the feature_index and threshold in each
                node
```

```
            Store the predicted class in an array pred_y
            Append majority in pred_y to vote_y
    end for
    end for
    return vote_y
```

4.6 Experimental Results and Performance Analysis

This section presents the experimental results and performance analysis of SVM and RF Model for dating ancient Kannada epigraphic scripts of varying periods.

4.6.1 SVM Model with Zone-based features

- **Experimental Dataset**

The Period Identification system is trained with ten instances of each unique character of ancient Kannada scripts, during the regime of different dynasties across varying period such as:- Ashoka, Satavahana, Kadamba, Badami Chalukya, Rashtrakuta, Kalyan Chalukya, Hoysala, Vijayanagar and Mysore Wodeyar dynasty. The Predictor or Era Identifier is tested on nearly 110 ancient Kannada epigraphical records of varying times.

- **Experimental Results**

The Period Identification system classifies the input Kannada epigraphs during the regime of different dynasties across varying period. In this section few sample experimental results of Era Prediction is illustrated in Figure 4.5 to Figure 4.11 showing Classification of the period for the given input epigraphical script, for periods Ashoka, Hoysala, Kadamba, Rashtrakuta, Kalyana Chalukya, Badami Chalukya and Mysore Wodeyar in order.

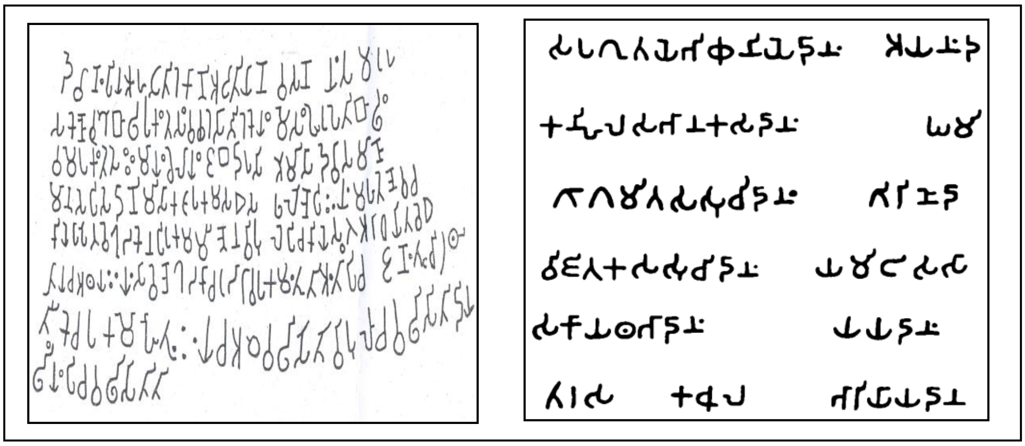


Figure 4.5: Sample Brahmi Scripts of Ashoka Period being Dated

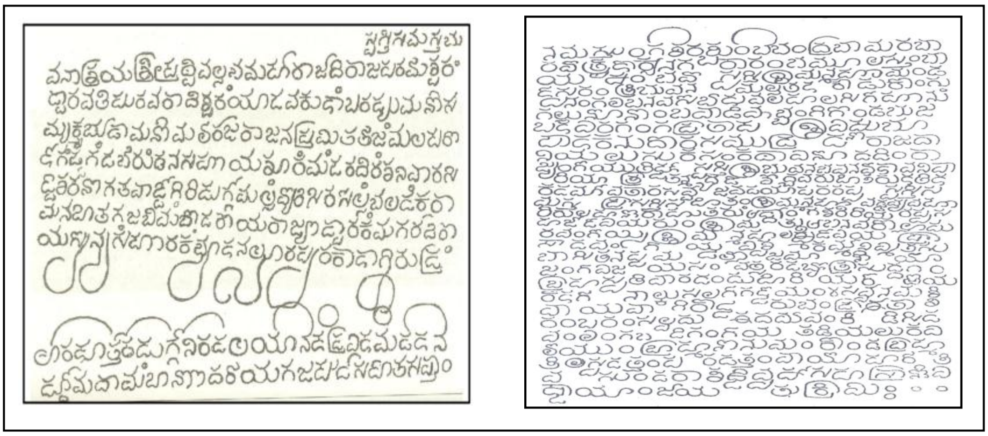


Figure 4.6: Sample Kannada Scripts of Hoysala Period being Dated

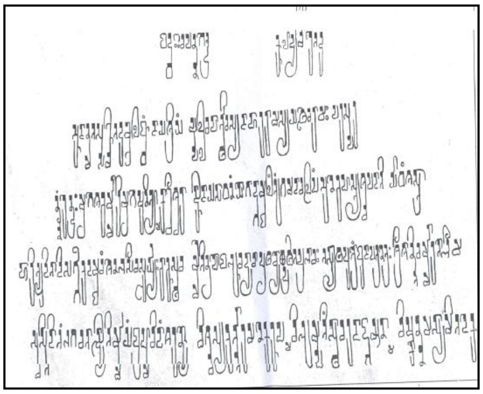


Figure 4.7: Sample Kannada Scripts of Kadamba Period being Dated

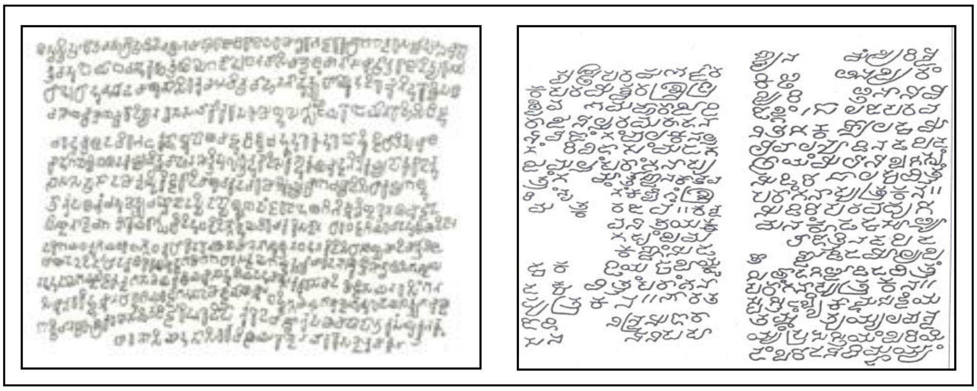


Figure 4.8: Sample Kannada Scripts of Rashtrakuta Period being Dated

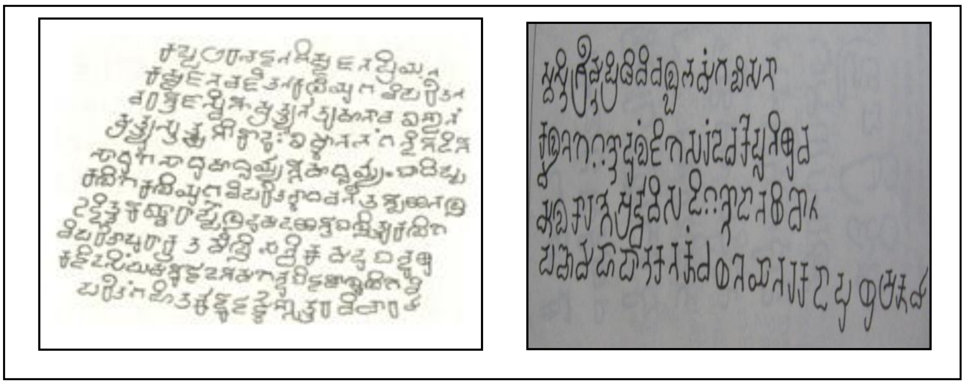


Figure 4.9: Sample Kannada Scripts of Kalyana Chalukya Period being Dated

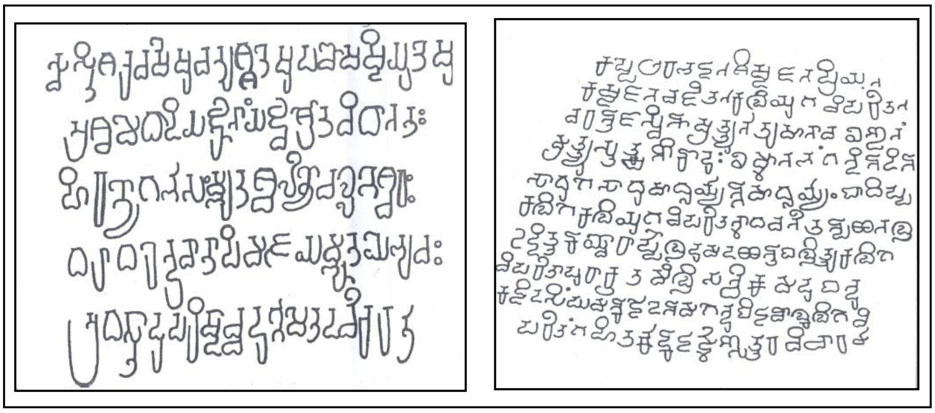


Figure 4.10: Sample Kannada Scripts of Badami Chalukya Period being Dated

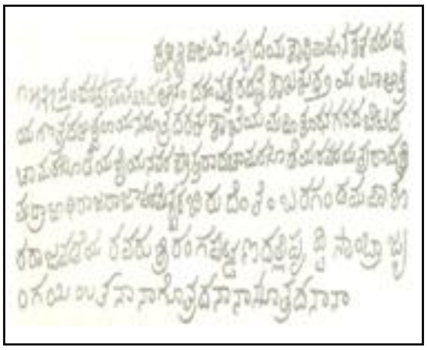


Figure 4.11: Sample Kannada Script of Mysore Wodeyar Period being Dated

- **Performance Analysis**

The proposed model for the prediction of the era is carried out by examining few characters, referred to as unique characters which uniquely characterizes each era in Kannada script of various periods. These unique characters of different periods form the training data and hence minimize the training time. It also reduces the search space during testing as only characters with unique features with respect to the era are considered for classification. The period identification using the concept of unique characters comparatively reduces the classification time. However, if the input script does not contain the unique characters then the model fails to identify the era of the epigraphic document. The Predictor or Era Identifier is tested on nearly 110 ancient Kannada epigraphic records from varying periods. The results are found to be satisfactory, with a average prediction rate of 80%.

4.6.2 RF Model with Central and Zernike features

This section presents experimental results of the RF model in the period identification and also discusses the performance characteristics of the designed RF Classifier by varying its parameters (number of trees and number of thresholds).

- **Experimental Dataset**

The dataset comprises of reconstructed images of ancient Kannada epigraphic scripts. The RF classifier is trained with all the base characters pertaining to different periods. 110 images belonging to Ashoka, Satavahana, Kadamba, Chalukya, Rastrakuta and Hoysala periods are used for testing the era predictor.

- **Experimental Results**

The designed RF model is tested on 110 epigraphic images belonging to Ashoka, Hoysala, Kadamba, Rashtrakuta, Kalyana Chalukya, Badami Chalukya and Mysore Wodeyar periods. The sample epigraphs dated are depicted in Figure 4.5 to 4.11

- **Analysis of Prediction rate**

A novel approach for predicting era using RF Classifier with Normalized Zernike Moments of degree 6 and Normalized Central Moments features is designed. The number of Classification Trees is varied and the prediction rates are observed. The results got are very encouraging, with average 85% prediction rate for the era identification on using a RF with 15 trees. 110 images of different eras are used for analyzing the prediction rate. All the epigraph images are tested with the threshold set to 10 and varying the number of trees and the result of percentage of correct predictions are tabulated in Table 4.1.

Table 4.1: Period Prediction rate (%) of Epigraphic images

Number of Trees	Prediction Rate
5	75.76
10	81.82
15	84.85

Figure 4.12 illustrates the above results showing the prediction rate of all the epigraph images for the threshold value 10 and a varying number of trees in RF classifier.

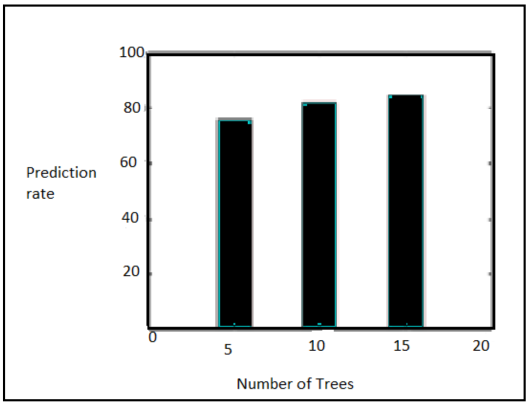


Figure 4.12: Period Prediction rate(%) for a threshold value of 10

It is observed that there is a linear increase of classification rate as the number of trees in the forest is increased.

4.7 Summary

The proposed work enables the user to classify the epigraph images of ancient Kannada scripts from different eras using SVM and RF model.

The SVM model for identifying the period of ancient Kannada epigraphs are found to be computationally inexpensive as the epigraphs are dated using only a few unique characters out of the entire character bank. The SVM classifier results are found to be satisfactory in determining the period of given input script. SVM works well even when the training dataset is minimal, as the classifier is trained with only a few distinct/unique characters of the era. The developed system fails to work on the epigraphical images without the existence of unique characters. Also it is observed that the statistical features such as pixel density, mean, variance, standard deviation, a fraction of the black to white pixels are primitive and are not invariant to image transformations such as translation, rotation and scaling. Hence, a novel approach of designing RF classifier for prediction of the period of the given epigraphical script, using statistical features such the Central Moments and Zernike moments invariant to transformations is proposed. Better prediction results are achieved using an ensemble classifier RF which is trained with the all the base characters of the era rather than only a few unique characters. This approach of dating epigraphs is computationally expensive, but gives better classification results.

Chapter 5

RECOGNITION OF EPIGRAPHICAL CHARACTERS AND APPROACHES

5.1 Premise

In Chapter 3 techniques for preprocessing and segmentation of epigraphical document images were discussed. Chapter 4 covered approaches for era identification of epigraphical records. Next step is the automatic decipherment of inscriptions which plays a significant role in knowing the cultural heritage and civilization that prevailed in ancient times. Hence in this chapter techniques are explored for the classification and recognition of epigraphical characters. Different models are discussed with variations of feature extraction techniques and classifiers for the recognition of epigraphical characters.

The chapter is organized as follows: The need for recognition of epigraphs is highlighted in Section 5.2. The proposed model for recognition of epigraphic documents is depicted in Section 5.3. Section 5.4 lists the approaches used for classification and recognition for epigraphic documents. In Section 5.5 to 5.10, the listed models for recognition of ancient text with experimental results and performance analysis is presented: - Zernike Features with SVM Classifier; Central and Zernike Moment features with RF Classifier; Zone-based and Gabor features with ANN; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier; SURF features with SVM, ANN and k-NN classifiers; and lastly, First-Order and Second-order statistical features with designed Fuzzy Classifier. Finally, Section 5.11 provides the summary of the chapter.

5.2 Significance of Epigraphic Character Recognition

The expert epigraphists decipher the text of ancient epigraphic scripts and translate them into regional languages. Also, it is observed that expert epigraphists who are capable of deciphering the inscriptions manually are few nowadays and they could become extinct in future. Modern readers find difficulty in reading the documents of ancient times. Hence, the automation of deciphering of the inscription is the need of the hour and is imperative. From the literature review it is evident that no substantiating work is done for deciphering epigraphical scripts.

The aim of the work here is to develop an automated system for classification and recognition of an ancient Kannada epigraphic text, whose period has been identified. Different methods are proposed to decipher text of ancient times with a combination of varying feature extraction techniques and classifiers. The methods are: Zernike Features with SVM Classifier; Central and Zernike Moment features with RF Classifier; Zone-based and Gabor features with ANN; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier; SURF features with SVM, ANN and k-NN classifiers and lastly First Order and Second order statistical features with designed Fuzzy Classifier for recognition of epigraphic characters.

5.3 Proposed Model for Epigraphic Character Recognition

The proposed model for the Epigraphical Character Recognition is shown in Figure 5.1 and involves the following components:

- **Preprocessing:** The input epigraphic image is preprocessed to remove noise.
- **Segmentation:** The noise-free epigraphs are segmented to obtain sampled characters.
- **Feature Extraction:** Essential features are extracted from the sampled characters and saved in a file during the training phase. During testing, the same set of features is extracted for test characters.
- **Database:** The database here represents the file used to save the extracted features. The feature vectors are used for training the classifier or for the recognition of characters during testing.
- **Classifier:** The Classifier is trained using the features stored in the file during training phase. It is also possible to save the trained Classifier for later use. The trained Classifier is used to classify the test characters during testing phase. The classified ancient characters are mapped to modern form and displayed.

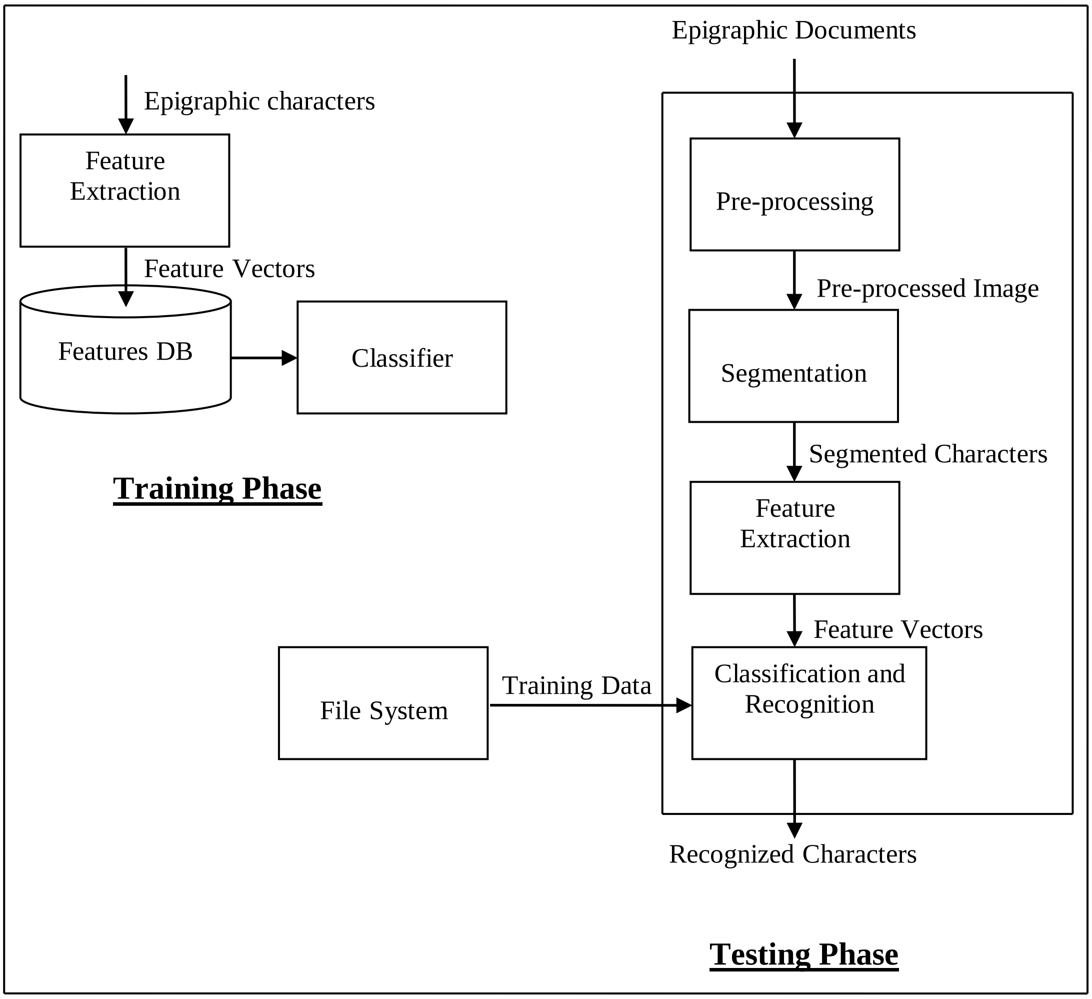


Figure 5.1: Proposed Model for Epigraphic Character Recognition with Training and Testing phases

- **Methodology for Recognition**

Algorithm: **RECOGNITION (Epigraph_Image)**

Input: Epigraphical document image

Output: Classified and Recognized characters in modern form

Method:

[Training Phase]

Step 1: Preprocess and Binarize the training epigraph images

Step 2: Segment the characters

Step 3: Extract the features of each of the characters

Step 4: Train the classifier using these features

[Testing Phase]

Step 5: Preprocess and Binarize the test epigraph image

Step 6: Segment the characters

Step 7: Extract the features of each of the test characters

Step 8: Classify each test character using the trained classifier

Step 9: Map the classified characters to modern form.

5.4 Methods for Recognition of Epigraphic Records

In the current work, the methods used to decipher text of ancient times with combination of varying feature extraction techniques and classifiers are: Zernike Features with SVM Classifier; Central and Zernike Moment features with RF Classifier; Zone-based and Gabor features with ANN classifier; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier; SURF features with SVM, ANN and k-NN classifiers and lastly First-order and Second-order Statistical features with Fuzzy Classifier for recognition of ancient epigraphic documents.

A survey of different shape feature extraction techniques is reported in [78]. The different image classification methods, and techniques for improving classification performance is reported in [79, 81].

➢ **Classification**

- **Support Vector Machine (SVM) Classifier**

Support Vector Machines (SVM) is a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyper-plane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier.

- **Artificial Neural Network (ANN) Classifier**

ANN usually called "neural network" (NN), is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.

- **k-Nearest Neighbor(k-NN) Classifier**

The k-Nearest Neighbor (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning- or lazy learning where the function is only approximated locally and all computation is deferred until classification. It can also be used for regression.

- **Naive Bayes Classifier**

A Naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without using any Bayesian methods.

5.5 Zernike Features with SVM Classifier

The work considers off line recognition of Epigraphical Kannada characters from three ancient eras Ashoka, Badami Chalukya and Mysore Wodeyar.

The preprocessed and segmented epigraphic characters from input epigraphic document image are fed to the feature extraction phase. The feature extraction method used is Zernike moments (discussed in Chapter 4). The advantages of using Zernike moments are that they are invariant to rotation, robust to noise and minor variations in shape and contain minimum information redundancy [78]. The features extracted for epigraphic characters are fed to the SVM classifier. The SVM classifier classifies the character of ancient age using support vectors and next the character is mapped to present Kannada form.

5.5.1 Methodology

The steps towards the classification and recognition of epigraphic characters are given here.

Algorithm: **Recognition (Epigraphic_Base_Characters)**

Input: Segmented base characters of ancient epigraphs

Output: Classified and Recognized characters

Step 1: Perform the following steps during training for segmented characters:

a: Compute Zernike features for base characters of training data

b: Compute the average feature value and store in the data base for later use.

c: SVM classifier is used to produce a model (based on the training data)

Step 2: Compute the Zernike features for each of the base characters during testing.

Step 3: SVM using support vectors from the training database, compares with the test character features and predicts the target values of the test data.

Step 4: Finally the classified characters are recognized, and mapped to the present Kannada form.

5.5.2 Experimental Results

This approach of Zernike features with SVM classifier demonstrates the recognition of base characters of ancient Kannada Script pertaining to three periods or dynasties – Ashoka, Badami Chalukya, and Mysore Wodeyars. The system is trained with all basic symbols of Ashoka, Badami Chalukya, and Mysore Wodeyars period. The recognizer has been tested on more than 250 samples of ancient Kannada epigraphic characters belonging to three different periods. The character recognition system successfully recognizes the base characters from three different periods and maps it to modern Kannada character. Figure 5.2 shows the recognition of test letter ‘ka’ from Ashoka period. The input ancient character is classified and recognized, and the character mapped to present Kannada form is displayed. Figures 5.3 and 5.4 show the recognition of letter ‘ja’ from Badami Chalukya era and letter ‘ou’ from Mysore Wodeyar era respectively.

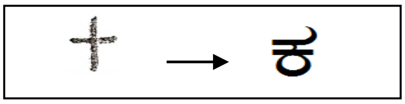


Figure 5.2: Recognition of letter ‘ka’ of Ashoka era using SVM Model

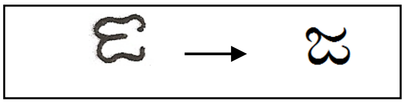


Figure 5.3: Recognition of character ‘ja’ from Badami Chalukya period using SVM

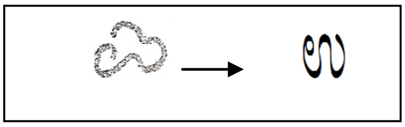
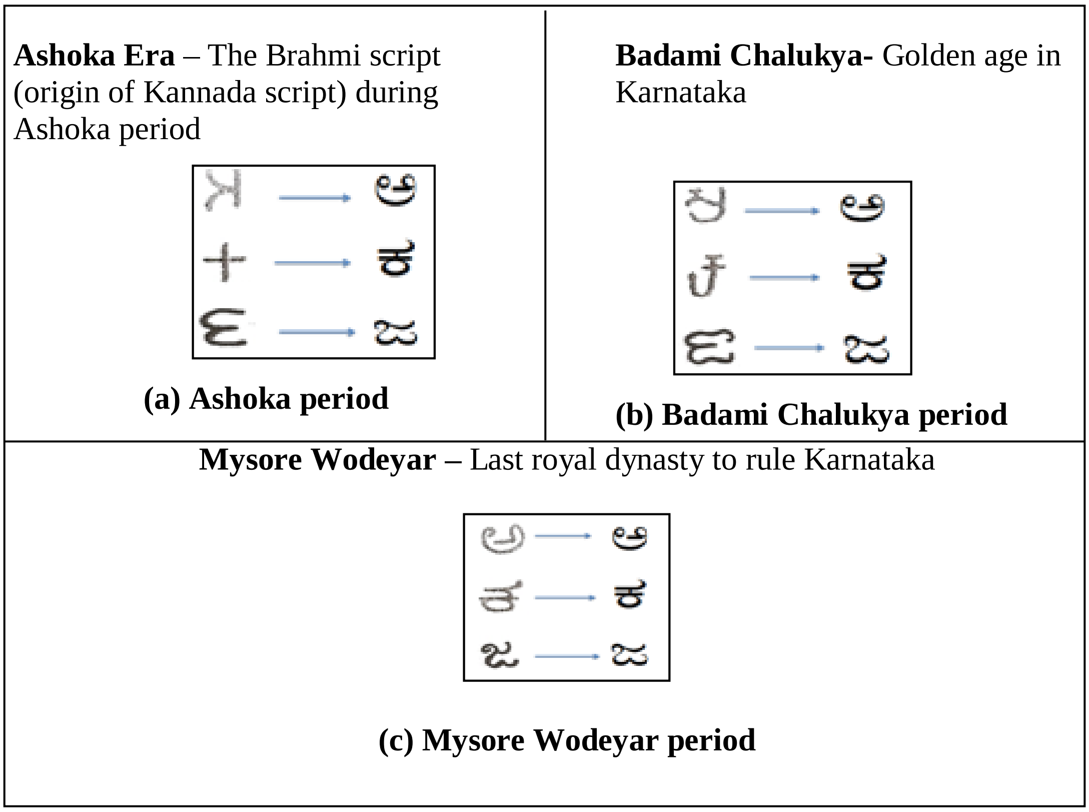


Figure 5.4: Recognition of character ‘ou’ from Mysore Wodeyar period using SVM



Ashoka Era – The Brahmi script (origin of Kannada script) during Ashoka period

(a) Ashoka period

Badami Chalukya- Golden age in Karnataka

(b) Badami Chalukya period

Mysore Wodeyar – Last royal dynasty to rule Karnataka

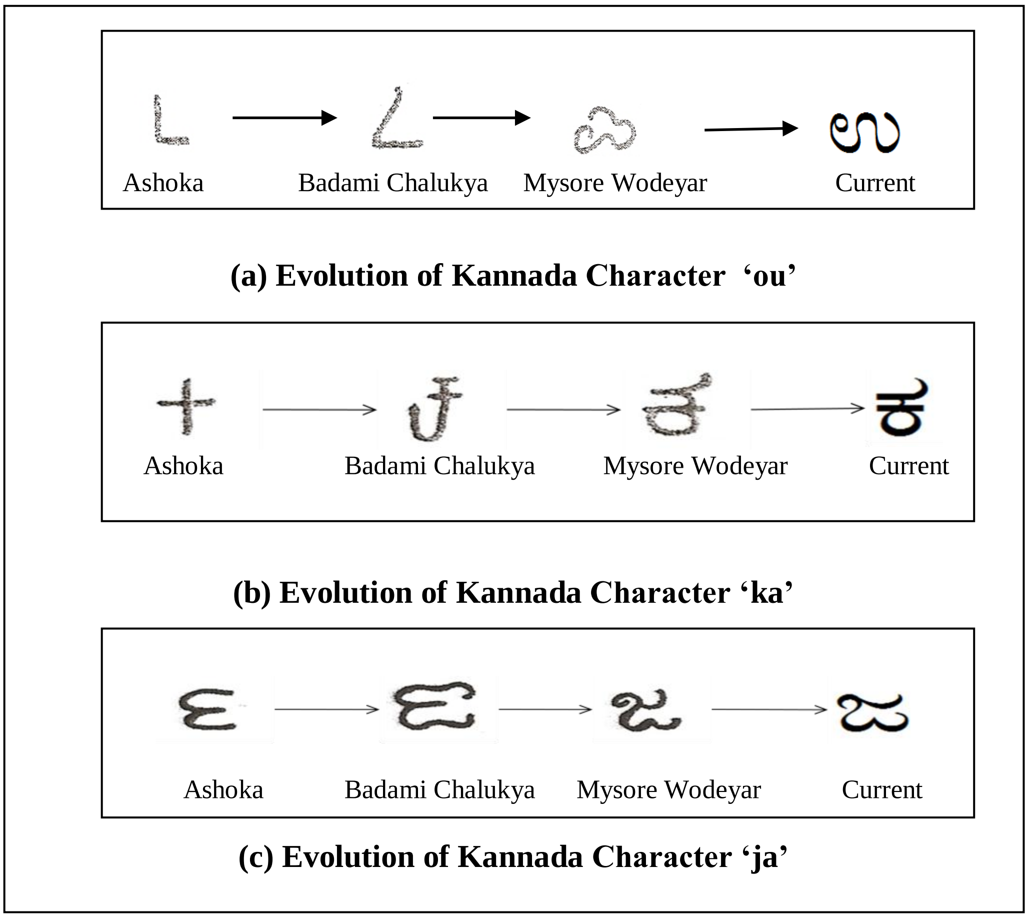
(c) Mysore Wodeyar period

Figure 5.5: Recognition of Sample Characters using Zernike features with SVM

Few other sample characters recognized and mapped to modern Kannada form of Ashoka, Badami Chalukya and Mysore Wodeyars period are depicted in Figure 5.5(a)-(c) respectively.

The proposed model recognizes base characters from three different eras Ashoka, Badami Chalukya and Mysore Wodeyar. Characters from these eras are also mapped to present Kannada Character set. Thus, recognition of such ancient characters gives the knowledge of how characters have evolved over generations and transformed to modern form. The Figures 5.6(a) - 5.6(c) show the evolution of few sample characters.

The recognizer has been tested on more than 250 samples of ancient Kannada epigraphic characters belonging to three different periods. The model achieves an average 90% recognition accuracy. The recognition accuracy rate of Ashoka era is 93%, Badami Chalukya is 90% and that of Mysore Wodeyar era is 88%.



(a) Evolution of Kannada Character ‘ou’

(b) Evolution of Kannada Character ‘ka’

(c) Evolution of Kannada Character ‘ja’

Figure 5.6: Evolution of Sample Ancient Kannada Characters

5.6 Central and Zernike Moment Features with RF Classifier

The RF Classifier was designed for dating ancient epigraphs as discussed in Chapter 4. This section covers the details of classification and recognition of Kannada epigraphical characters using the earlier designed RF classifier.

Normalized Central Moments and Zernike Moments are extracted from the segmented characters and used as the feature vectors for classification. Random Forest is used as the classifier, which is an ensemble of classification trees, and each tree votes for a class and the output class is the majority of the votes [77, 78]. Thus, all the characters in the image are classified. Finally the classified ancient characters are mapped to characters of modern form.

5.6.1 Methodology

***Algorithm*: RECOGNITION (Epigraph_Image)**

Input: Segmented epigraphic characters.

Output: Classified and Recognized characters.

Method:

Step 1: [Feature Extraction]: The Normalized Central Moments and Normalized Zernike Moments are computed, and the computed feature vectors are written to a file.

Step 2: [Random Forest Classification]

- **a. [Load Text]:** Get the feature vectors from the text file and save it in two arrays, one consisting of the classes and the other consisting of feature vectors of the corresponding classes.
- **b. [Fit Forest]:** Train the trees in the RF which can be used to classify the ancient Kannada characters.
- **c. [Fit Tree]:** A random subset of the training data from the step Fit Forest is taken as input and a single Classification Tree for the given subset of data is made.
- **d. [Get Gini Impurity]:** Determine the impurity index of a subset of classes and corresponding data for the node so that it can find the best split and the best threshold value for that feature.
- **e. [Classification]:** Predict the class of the test characters considering the data consisting of feature vectors, using the trained RF Classifier.

Step 3: [Recognition]: Map the classified ancient characters into modern form.

5.6.2 Experimental Results and Performance Analysis

The experimental results and analysis of the designed RF for classifying ancient Kannada Epigraphical characters are discussed here. The system is tested on base characters belonging to Ashoka, Satavahana and Kadamba dynasties. For each dynasty, 105 samples with 35 base characters are considered. Two-thirds of the data is used for training and the remaining one-third is taken for testing the classifier.

➢ **Performance Characteristics of RF Classifier**

- **Evaluation Metrics**

The metrics used to evaluate the proposed model are:

- **Classification rate:** This metric given by Equation 5.1 is used to determine the accuracy of the Classifier, which is defined as the number of correct classifications out of the total number of samples considered.

$$\text{Classification rate } = \frac{\text{Number of correctly classified characters}}{\text{Total number of characters in the data}} \qquad (5.1)$$

- **Training time**: This metric measures the time taken to train the Classifier.
- **Classification time:** The classification time is the time taken to predict the class labels for the given set of inputs.

- **Classification Rate of RF on the Characters from Satavahana Period**

The accuracy of RF in classifying characters from trained data set of Satavahana period for the threshold value 10 and a varying number of trees are tabulated in Table 5.1. The plot in Figure 5.7 shows the results of the same on trained data.

Table 5.1: Classification Rates (%) of RF Model for Trained Data

	Number of Trees			
Thresholds		**10**	**20**	**30**
	10	47.69	81.54	90.77
	20	50.77	69.23	84.62
	30	49.23	73.85	87.69

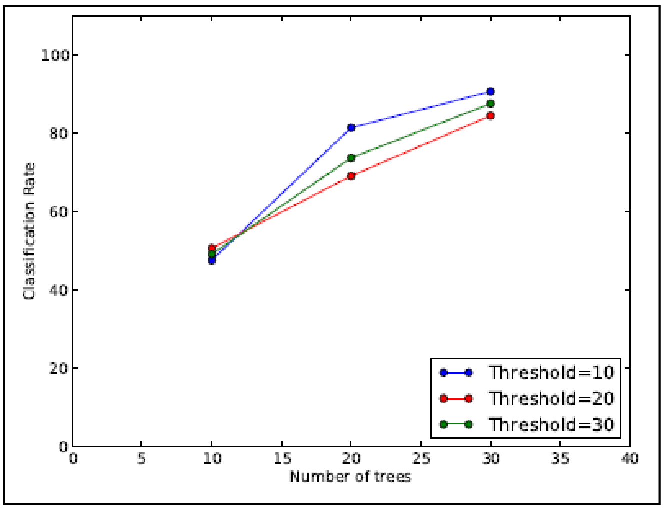


Figure 5.7: Classification Rates of RF with different Parameters for Trained data

The Classification rate for test characters is tabulated in Table 5.2 and shown in Figure 5.8.

Table 5.2: Classification Rates (%) of RF Model for Test Data

	Number of Trees			
Thresholds		**10**	**20**	**30**
	10	43.53	52.35	67.06
	20	44.53	52.35	70.00
	30	58.25	53.35	61.18

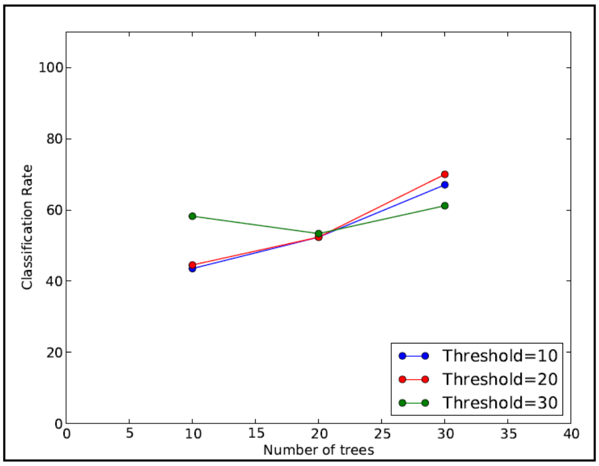


Figure 5.8: Classification Rates of RF with different Parameters for Test data

- **Training and Testing time of RF on the characters from Satavahana period**

The time (seconds) for training characters from Satavahana period are tabulated in Table 5.3 and plotted in Figure 5.9 respectively. As the number of trees in the forest increases, the time taken for training also increases proportionately.

Table 5.3: Training Time (seconds) of Random Forest Model

	Number of Trees			
Thresholds		**10**	**20**	**30**
	10	3.53	7.42	11.62
	20	6.62	11.99	19.07
	30	7.78	15.84	22.64

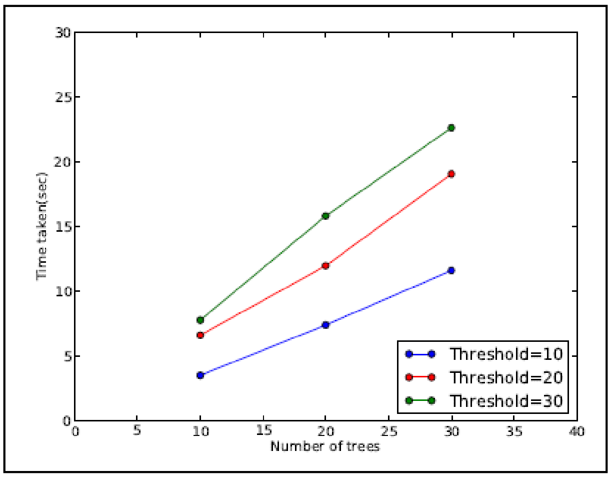


Figure 5.9: Training Time for RF with different Parameters

The classification times (in seconds) for new characters are tabulated in Table 5.4 and the plot for the same for different parameters is shown in Figure 5.10

Table 5.4: Classification Time Taken In Seconds for RF Model

	Number of Trees			
Thresholds		**10**	**20**	**30**
	10	0.0183	0.0361	0.5436
	20	0.0187	0.0367	0.0553
	30	0.0181	0.0365	0.5343

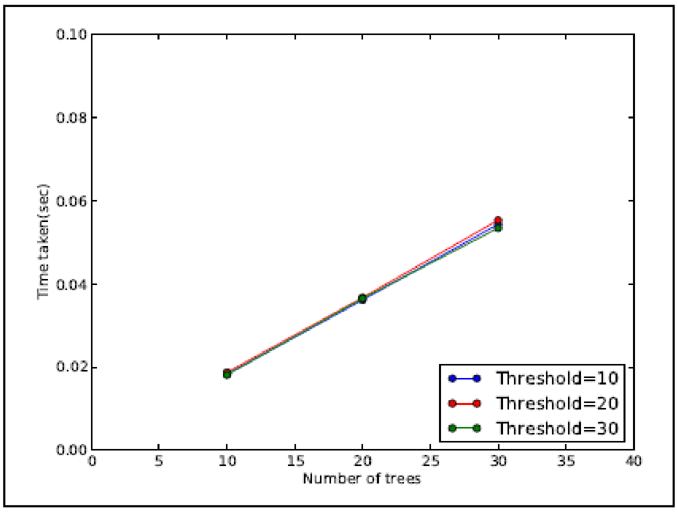


Figure 5.10: Classification Time for RF with different parameters

The training time taken by the RF Classifier using samples with 35 base characters from Satavahana period were 3.5, 7.4 and 11.6 seconds for RF with 10, 20 and 30 classification trees, respectively, for a threshold of 10. When the number of thresholds was increased to 20, the times taken to train were 6.6, 12 and 19 seconds for RF with 10, 20 and 30 classification trees, respectively. When the number of thresholds was 30, the training times were 7.8, 15.8 and 22.6 seconds for RF with 10, 20 and 30 classification trees, respectively. But the classification rate changed only in the range of 4%-10%. Hence, fixing the number of thresholds at 10 would be a good tradeoff between training time and classification rate.

The following inferences are drawn from the performance analysis:

- The accuracy in classification of the trained data is at least 1.2 times greater than the classification rate of new characters for any classifier.
- There is a linear increase of classification rate as the number of trees in the forest is increased, but no significant changes when the number of thresholds is increased.
- The training time is directly proportional to the number of classification trees and the number of thresholds.
- The classification time is directly proportional to the number of classification trees. It is not dependant on the number of thresholds since it is used only when growing the trees.
- The training time of the RF classifier is about 200 times more than the classification time. This is because most of the time is spent for the calculation of Gini index during training. Classification involves only a comparison at each node till it reaches the leaf.

5.7 Zone-based and Gabor features with ANN Classifier

In this section recognition of epigraphic characters using Zone-based and Gabor features with Neural network classifier is discussed.

5.7.1 Methodology

This work includes steps: Pre-Processing Segmentation, Feature Extraction, Recognition and Post-Processing.

***Algorithm*: RECOGNITION (Epigraph_Image)**

Input: Scanned Epigraphic document.

Output: Classified and Recognized characters.

Method:

Step 1: Input scanned ancient Kannada epigraph to the recognizer.

Step 2: [Preprocess]: Preprocess and **S**egment to extract individual characters.

Step 3: [Feature Extraction]: Extract Zone-based and Gabor features for segmented characters and store in the feature vector.

Step 4: [Training]: Train Artificial Neural Network with feature vectors of the sampled train characters.

Step 5: [Classification]: Classify the segmented characters from test images using the trained ANN classifier

Step 6: [Mapping]: The classified ancient characters are mapped to modern form.

5.7.2 Related Theory and Mathematical Background

- **Zone-based feature extraction**

In character recognition, zoning [78] is used to extract topological information from patterns. The segmented image is divided into ‘n’ zones and from each zone statistical features like number of horizontal/vertical/diagonal lines, length of horizontal/vertical/diagonal lines, the total number of intersection points are extracted.

- **Gabor Feature extraction**

Gabor filters act very similarly to mammalian visual cortical cells so they extract features from different orientation and different scales [80]. The filters have been shown to possess optimal localization properties in both spatial and frequency domain and thus are well suited for texture segmentation problems. Gabor filters have been used in many applications, such as texture segmentation, target detection, document analysis and edge detection. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave.

The impulse response of Gabor filter is defined by a sinusoidal wave (a plane wave for 2D Gabor filters) multiplied by a Gaussian function. Because of the multiplication-convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function. The filter has a real and an imaginary component representing orthogonal directions. The two components may be formed into a complex number or used individually.

Complex

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \qquad (5.2)$$

Real

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi\frac{x'}{\lambda} + \psi\right) \qquad (5.3)$$

Imaginary

$$(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \sin\left(2\pi\frac{x'}{\lambda} + \psi\right) \qquad (5.4)$$

where

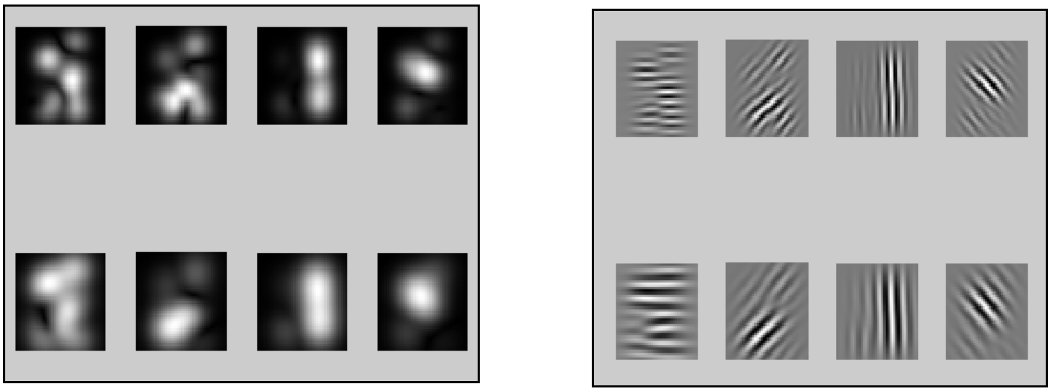
$$x' = x \cos\theta + y \sin\theta$$

and

$$y' = -x \sin\theta + y \cos\theta$$

In these equations, λ represents the wavelength of the sinusoidal factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function, ψ is the phase offset, σ is the standard deviation of the Gaussian envelope and γ is the spatial aspect ratio- and specifies the ellipticity of the support of the Gabor function.

Figure 5.11 shows Gabor Filtered images of a sampled character at two scales and four orientations.



(a)Magnitudes of Gabor Filter **(b)Real parts of Gabor Filter**

Figure 5.11: Gabor Filtered Images of a Sampled Character

5.7.3 Detailed Description and Algorithms

- **Feature Extraction**

The purpose of this phase is to extract the features of the epigraphical character images.

- **Gabor features**

This function uses Gabor filter methods to extract features from the character image.

***Algorithm*:** Gabor_Feature_Extract (Segmented Character)

Input: Segmented Characters

Output: Feature Vector of characters

Method

Step 1: Compute the orientation

Step 2: Compute the gabor filter bank

Step 3: Convolve it using the conv2 function.

Step 4: Down sample the image by factors of the size of image

Step 5: Store the resultant value in a feature vector

- **Zonal features**

This function extracts Zone-based features from the character image.

Algorithm: Zone_Based_Features (Segmented Character)

Input: Segmented Characters

Output: Feature Vector of characters

Method

Step 1: Input image is divided into 9 zones of equal size

zone_{ij} = image(1:zone_height,1:zone_width);

Step 2: From each zone following features are extracted.

The number of horizontal lines, total length of horizontal lines, number of right diagonal lines, total length of right diagonal lines, number of vertical lines, total length of vertical lines, number of left diagonal lines, total length of left diagonal lines and number of intersection points

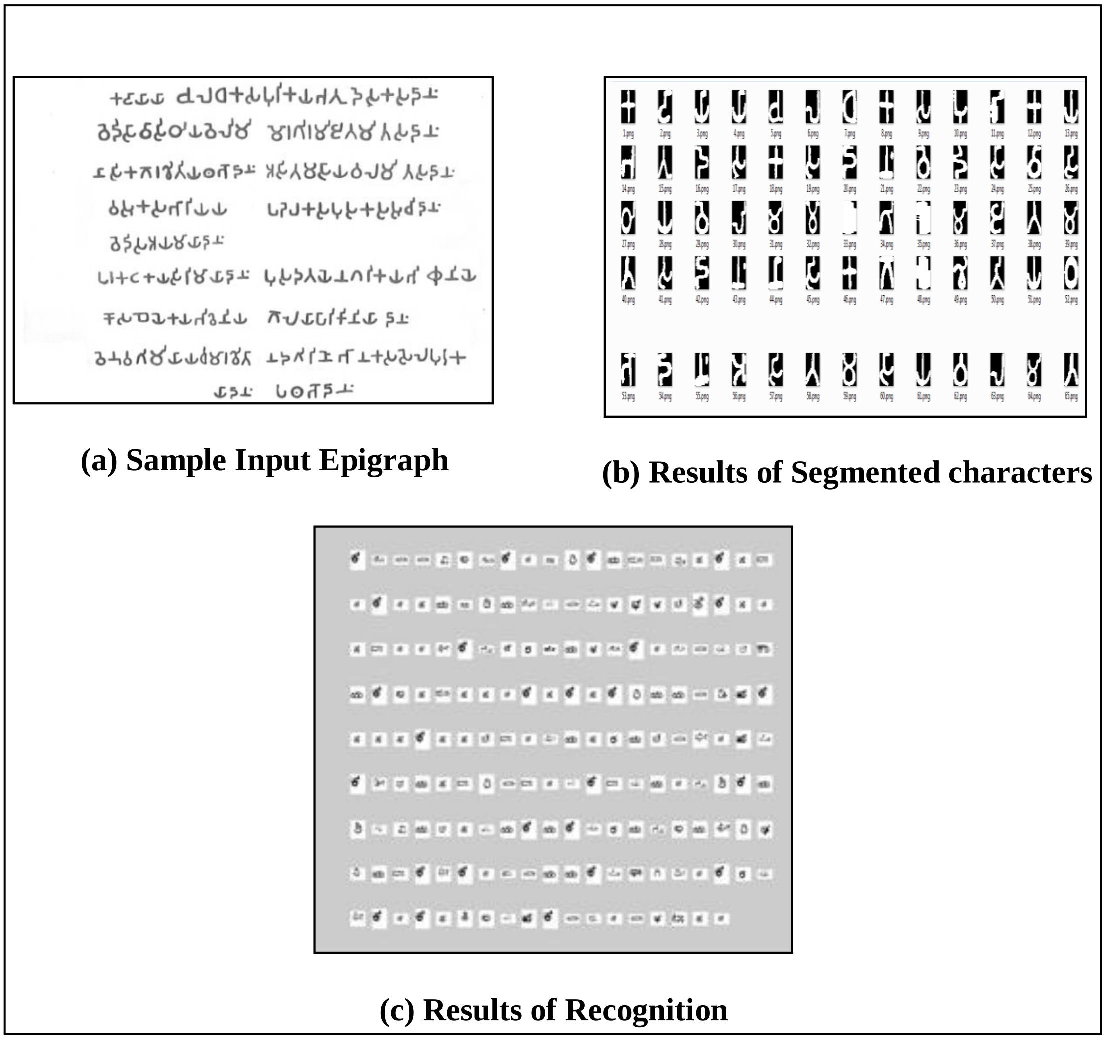
Step 3: Extracted features are stored in the new feature vector.

- **Mapping**

Final mapping of each classified character to the modern Kannada character is done. Class label returned by classifier is used to match with the modern character database and that character is displayed on the screen.

5.7.4 Experimental Results and Analysis

Figure 5.12 illustrates the results of Classification and Recognition of input epigraph of Ashoka period.



(a) Sample Input Epigraph

(b) Results of Segmented characters

(c) Results of Recognition

Figure 5.12: Classification and Recognition of Epigraph using Gabor Features

- **Performance Analysis**

- **Training: Ashokan Brahmi script**

The training database of Ashokan Brahmi script contains 8 vowels, 33 consonants which give rise to 264 different compound characters. So there are 272 different characters hence 272 class labels.

Four instances of each character are used, so this gives raise to (264+8=272) x 4 = 1088 characters which forms input for training and target vectors indicates the class to which each of the input character belongs.

- **Testing: Ashokan Brahmi script**

The model is tested with 100 epigraphic images of Ashokan dynasty and obtains an average recognition accuracy of 80.2%.

- **Training: Hoysala script**

The training database of Hoysala script contains 11 vowels and 36 consonants which give rise to 396 different compound characters. So there are total 407 different characters hence 407 class labels.

Four instances of each character were used, so this gives raise to (396+11=407) x 4 = 1628 characters which form input for training and target vectors indicates the class to which each of the input character belongs.

- **Testing: Hoysala script**

The model is tested with 50 epigraphic images of Hoysala dynasty and obtains an average recognition accuracy of 75.6%.

5.8 Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier

5.8.1 Methodology

The approach takes an image of an epigraph pertaining to ancient Kannada script as its input. The image is preprocessed to remove noise. The Preprocessed image is segmented using Canny edge detection which extracts the edges of ancient script characters. Close character contours are detected from the edges, based on that characters are segmented and stored in the database. General Fourier features are extracted from the segmented characters and based on this the Size and Scale Invariant Fourier features are extracted and used as the feature vectors for classification. Next for classification four classifiers Support Vector Machine (SVM), Artificial Neural Network (ANN), K- Nearest Neighbor (k-NN), Naive Bayes (NB) classifiers [79] are used. These classifiers are trained with different instances of characters during the training phase and while testing categorizes the ancient characters in the test image. Finally, from the predicted class label ancient character is mapped to the modern Kannada character.

5.8.2 Related Theory and Background

Fourier features a$_n$, b$_n$, c$_n$, and d$_n$ are extracted from close character contours. From these general Fourier features, scale and rotation invariant features are extracted [78].

- **General Fourier features**

Fourier features can be extracted from close character contours. a$_n$, b$_n$, c$_n$, and d$_n$ are the extracted features and given by Equations 5.5 to 5.8 respectively.

$$a_n = \frac{T}{2n^2\pi^2}\sum_{i=1}^{m}\frac{\Delta x_i}{\Delta t_i}[\cos\phi_i - \cos\phi_{i-1}] \tag{5.5}$$

$$b_n = \frac{T}{2n^2\pi^2}\sum_{i=1}^{m}\frac{\Delta x_i}{\Delta t_i}[\sin\phi_i - \sin\phi_{i-1}] \tag{5.6}$$

$$c_n = \frac{T}{2n^2\pi^2}\sum_{i=1}^{m}\frac{\Delta y_i}{\Delta t_i}[\cos\phi_i - \cos\phi_{i-1}] \tag{5.7}$$

$$d_n = \frac{T}{2n^2\pi^2}\sum_{i=1}^{m}\frac{\Delta y_i}{\Delta t_i}[\sin\phi_i - \sin\phi_{i-1}] \tag{5.8}$$

where $\phi_i = {}^{2n\pi t_i}\!/_{T}$, $\Delta x_i = x_i - x_{i-1}$, $\Delta y_i = y_i - y_{i-1}$, $\Delta t_i = \sqrt{\Delta x^2 + \Delta y^2}$,

$T = t_m = \sum_{j=1}^{m}\Delta t_j$, $t_i = \sum_{j=1}^{i}\Delta t_j$ and m is the number of pixels along the boundary.

- **Rotation invariant Fourier features**

To obtain the features that are independent of the particular starting point, it is required to calculate the phase shift from the first major axis as in Equation 5.9

$$\phi_1 = \frac{1}{2}\tan^{-1}\frac{2\left(a_1b_1 + c_1d\ \right)}{\sqrt{a_1^{\,2} - b_1^{\,2} + c_1^{\,2} - d_1^{\,2}}} \tag{5.9}$$

Then, the coefficients can be rotated to achieve a zero phase shift as given by Equation 5.10

$$\begin{bmatrix} a_n^* b_n^* \\ c_n^* d_n^* \end{bmatrix} = \begin{bmatrix} a_n\, b_n \\ c_n\, d_n \end{bmatrix} \begin{bmatrix} \cos n\phi_1 - \sin n\phi_1 \\ \sin n\phi_1 \cos n\phi_1 \end{bmatrix} \tag{5.10}$$

Now to obtain rotation invariant description, the rotation of the semi-major axis can be found by Equation 5.11:

$$\psi_1 = \tan^{-1} \frac{c_1^*}{a_1^*} \tag{5.11}$$

Now using Equation 5.12 features can be obtained.

$$\begin{bmatrix} a_n^{**} b_n^{**} \\ c_n^{**} d_n^{**} \end{bmatrix} = \begin{bmatrix} \cos\psi_1 & \sin\psi_1 \\ -\sin\psi_1 & \cos\psi_1 \end{bmatrix} \begin{bmatrix} a_n^* b_n^* \\ c_n^* d_n^* \end{bmatrix} \tag{5.12}$$

- **Scale invariant Fourier features**

  To obtain Scale invariant features the coefficients can be divided by the magnitude, E, of the semi-major axis, given by Equation 5.13:

$$E = \sqrt{a_1^{*2} + c_1^{*2}} = a_1^{**} \tag{5.13}$$

5.8.3 Detailed Description and Algorithm

The steps involved in classification and recognition of epigraphic characters using Fourier features with SVM, ANN, k-NN, NB classifiers are given in this section.

Algorithm: RECOGNISE (Epigraph)

Input: Ancient Kannada Epigraph

Output: Modern Kannada characters

Method:

Step 1: Read the epigraph image.

Step 2: Preprocess epigraph

Step 3: Segment epigraph and store segmented characters.

Step 4: Extract Fourier features for segmented characters.

Step 5: Train Classifiers SVM, ANN, k-NN and NB using these features.

Step 6: Extract features of segmented test character.

Step 7: Classify the segmented character of the ancient period.

Step 8: Classified ancient character is mapped to the modern Kannada form.

Step 9: Return modern Kannada character.

- **Fourier Feature Extraction**

General Fourier features are extracted from close character contours. a_n, b_n, c_n, and d_n are the extracted features.

Algorithm: Fourier_Features (Character Contour, x, y)

Input: Segmented character image

Output: General Fourier features a_n, b_n, c_n, d_n.

Method:

Step 1: Initialize Parameters

Step 2: Compute Fourier features

Step 3: Store Fourier features

End Method

- **Classification**

  - **SVM Classifier**

This model classifies the segmented character and predicts its class label. It consists of the following steps:

Step 1: Set up the training data

Step 2: Set up the training classes

Step 3: Set up SVM's parameters

Kernel Type = LINEAR, SVM Type = C_SVC, Termination Criteria

Step 4: Train the SVM

Step 5: Classification of characters using SVM

- **ANN Classifier**

This model classifies the segmented character and predicts its class label. It consists of the following steps:

Step 1: Set up the training data

Step 2: Set up the training class

Step 3: Set up ANN's parameters

Parameters of the MLP training algorithm are set.

***term_crit*:** Termination criteria of the training algorithm.

***train_method*:**Indicates the training method of the MLP - back-propagation.

Step 4: Train the ANN

ANN model is built with MLP network and Activation_function SIGMOID

Step 5: Classification of characters using trained ANN

- **k-NN Classifier**

The k-nearest neighbors (k-NN) is a method for classifying objects based on closest training examples in the feature space. Here it classifies the segmented character and predicts its class label. It consists of the following steps:

Step 1: Set up the training data

Step 2: Set up the training classes

Step 3: Set up k-NN's parameters

Finds the neighbors

samples – Input samples stored by rows.

k – Number of nearest neighbors used.

results –results of prediction (classification) for each input sample.

Neighbor Responses –Optional output values for corresponding neighbors.

Step 4: Train the k-NN

Step 5: Classification of characters by k-NN

- **Naive Bayes Classifier**

A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. This classifies the segmented character and predicts its class label. It consists of the following steps:

Step 1: Set up the training data

Step 2: Set up the training classes

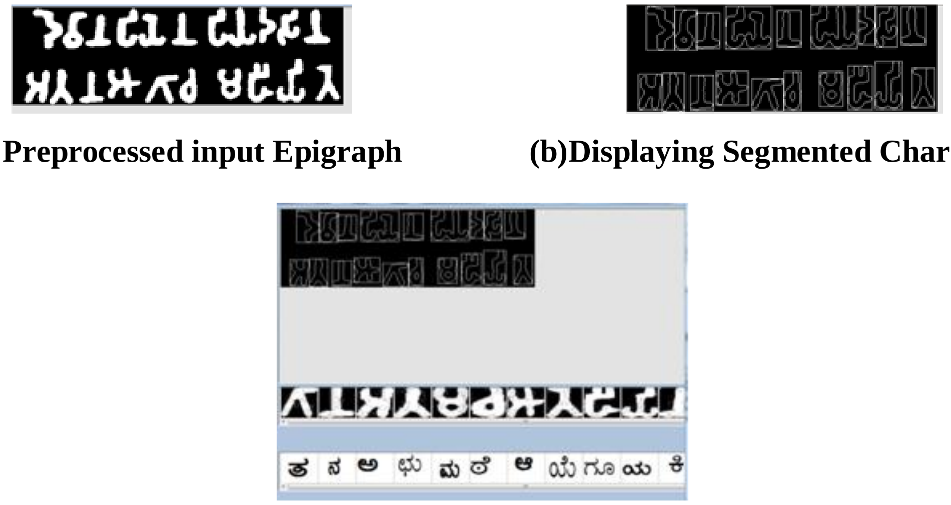
Step 3: Train the Naive Bayes

Step 4: Classification of characters by Naive Bayes

5.8.4 Experimental Results and Analysis

The system designed recognizes and converts Ashokan Brahmi script and Hoysala script into modern Kannada form. The system is trained with characters of Ashoka period and characters of Hoysala period with 4 different instances of each. The trained OCR system is tested on 50 epigraphs of ancient times.

Figure 5.13 shows the results of Preprocessing, Segmentation and recognition.



(a) Preprocessed input Epigraph

(b)Displaying Segmented Characters

(c) Results of Recognition using Fourier features

Figure 5.13: Results of Recognition using Fourier features

The performance characteristics of the Classifiers is obtained and observed that with Ashokan Brahmi script the system performed well with recognition accuracy of 83.60%, 76.80%, 49.20%, 64.80% . For Hoysala script, recognition accuracy of 80.50%, 71.50%, 48.50%, 62.50% with `SVM, ANN, k-NN and NB classifiers respectively is obtained.

5.9 SURF Features with SVM, ANN and k-NN classifiers

5.9.1 Methodology

The aim of this approach is to use multiple classifiers that recognize ancient Kannada characters and maps them to modern Kannada. The approach accepts ancient Kannada epigraph from Ashoka period as input. The input image is binarized by Adaptive thresholding and noise is removed by applying a combination of three filters namely Median, Bilateral and Gaussian Filters. Furthermore, ancient text is made prominent by applying erode and dilate morphological operations. The resultant image is passed to the Segmentation stage where Bounding Box and Contour Detection Algorithm are used to segment individual characters including the vattaksharas (compound characters). These segmented characters are then passed to the Feature extraction stage that makes use of SURF technique to create the feature vectors. Classification is the final stage shown in Figure 5.14, which is carried out in two phases namely Training and Testing. In order to train the Classifiers, the feature vectors are passed to the classifiers namely SVM, k-NN and ANN in the training Phase. In the testing phase, the feature vectors are passed to a trained classifier to recognize the ancient character. A combination of three classifiers namely, SVM, k-NN and ANN are used to achieve better accuracy. The recognized character is subsequently mapped to its modern equivalent. In this way, a document in ancient Kannada can be translated to modern Kannada.

5.9.2 Related Theory and Mathematical Background

- **Feature Extraction**

SURF or Speeded up Robust Features is a scale- and rotation-invariant interest point detector and descriptor [82]. SURF is a detector and high-performance descriptor points of interest in an image where the image is transformed into coordinates, using a technique called multi-resolution. It approximates or even outperforms previously proposed schemes with respect to repeatability, distinctiveness, and robustness, yet can be computed and compared much faster. The mathematical representation is given by Equation 5.14,

$$H(p,\sigma) = \begin{pmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{yx}(p,\sigma) & L_{yy}(p,\sigma) \end{pmatrix} \tag{5.14}$$

where $L_{xx}(p,\sigma)$ is the convolution of second order derivative $\frac{\partial x}{\partial x^2 g(\sigma)}$ with the image in the point x, y similarly with $L_{xy}(p,\sigma)$ and $L_{yy}(p,\sigma)$.

The SURF algorithm is based on the SIFT predecessor. This is achieved by

- Relying on integral images for image convolutions
- Building on the strengths of the leading existing detectors and descriptors (using a Hessian matrix-based measure for the detector, and a distribution-based descriptor)
- Simplifying these methods to the essential

This leads to a combination of novel detection, description, and matching steps. The detector is based on the Hessian matrix, but uses a very basic approximation, just as DoG is a very basic Laplacian-based detector [82]. It relies on integral images to reduce the computation time and therefore it is called the 'Fast-Hessian' detector. The descriptor, on the other hand, describes a distribution of Haar-wavelet responses within the interest point neighbour hood. The integral images are exploited for speed. Moreover, only 64 dimensions are used, reducing the time for feature computation and matching, and increasing simultaneously the robustness. A new indexing step based on the sign of the Laplacian is presented, which increases not only the matching speed- but also the robustness of the descriptor.

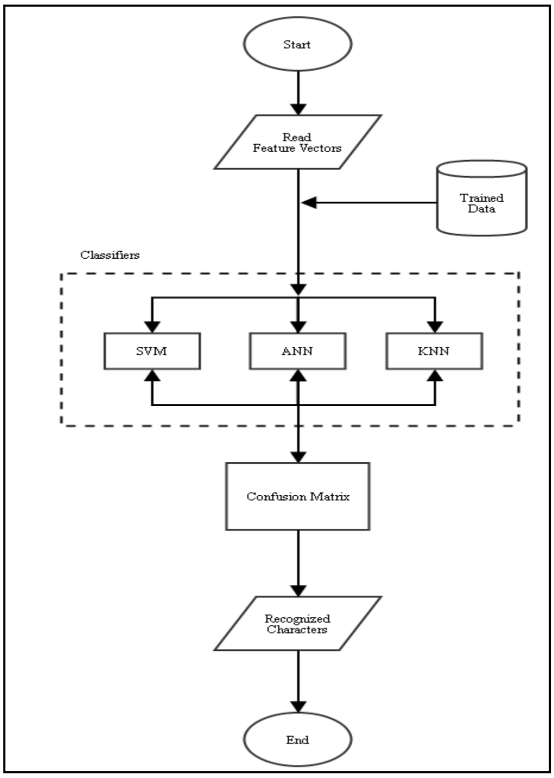


Figure 5.14: Model for Classification and Recognition using Multiple Classifier

5.9.3 Detailed Description and Algorithm

- **Feature Extraction**

In this stage, for each character segmented image, a hessian threshold is calculated that leads to detecting key points of the character to form Feature Vectors. Algorithm for SURF Feature Extraction technique is as follows:

Algorithm: SURF_ Feature_Extract (Epigraphic character)

Input: Individual Character Segment

Output: Feature Vectors

Functionality: Extracts the key-points from the image and creates the feature vectors.

Step 1: Input the necessary segmented character

Step 2: Set the Hessian Threshold to 450.

Step 3: Choose the SURF descriptor size to be 64 dimensions.

Step 4: Detect the SURF features.

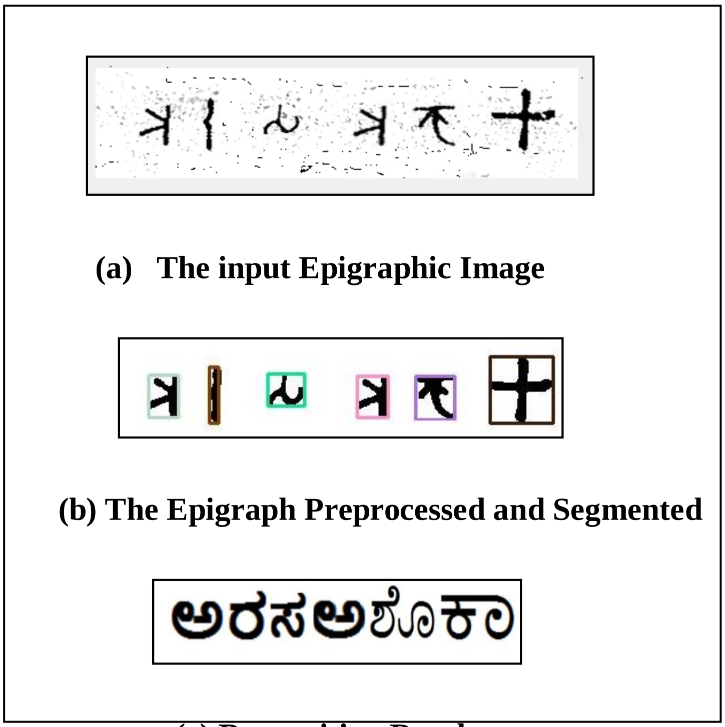
Step 5: Construct the feature descriptor for the detected features.

- **Classification and Recognition**

The SURF features are extracted for the test characters and classified using the classifiers- SVM, ANN and k-NN. For the efficient recognition of characters, the output of these multiple classifiers is passed to the Confusion Matrix. The Confusion Matrix evaluates the result of the classifiers and provides the final character label. The classified character is next mapped to modern form.

5.9.4 Experimental Results and Analysis

This OCR recognizes the ancient scriptures of Ashoka period using SVM, ANN and k-NN. Figure 5.15 represents the input epigraphic image. The Figure 5.16 shows the results of Preprocessing and Segmentation of the input epigraph. The output of recognition is shown in Figure 5.17. The system performs well with better recognition rate when multiple classifiers namely SVM, ANN and k-NN are used along with the confusion matrix in order to resolve the errors which arise during the character recognition. Thus, a combination of classifiers in recognizing characters gives a higher accuracy than using individual classifiers. The classifiers SVM, k-NN and ANN when tested on randomly chosen 90 characters achieve a recognition accuracy of 85%, 85% and 80%, but when the classifiers are combined the recognition accuracy increases to 95% thereby experimentally demonstrating that a combination of classifiers achieves 5-10% higher accuracy. However this may increase in time complexity due to the presence of more than one classifier. Care needs to be taken to decrease the time complexity and improve the recognition accuracy. Hence, the confusion matrix is introduced to create this win-win situation. Another approach is to execute the classification stage in parallel so as to reduce the time complexity.



(a) The input Epigraphic Image

(b) The Epigraph Preprocessed and Segmented

(c) Recognition Results

Figure 5.15: The Results of Recognition using SURF features

5.10 First-order and Second-order Statistical features with Fuzzy Classifier

5.10.1 Methodology

This approach includes phases: Segmentation, Feature Extraction and Classification in transforming input ancient Kannada epigraph into present Kannada form. As an initial step, Nearest Neighbor algorithm is used to segment characters from the given ancient Kannada input epigraph. Statistical features Mean, Variance, Standard Deviation, Skewness, Kurtosis, Entropy, Smoothness, Coarseness and other Histogram features are extracted from segmented characters. Next, Mamdani based fuzzy classifier is used to recognize the characters and transform into modern Kannada form.

5.10.2 Related Theory and Mathematical Background

This section discusses the related theory and fundamentals with the relevant mathematical background for the approaches used in the current work.

- **Feature Extraction**

Texture patterns are analyzed using statistical approach and the spatial relationship between the pixel values. Texture analysis plays an important role in pattern analysis and recognition. It determines the smoothness and coarseness of image features. Texture feature extraction has two approaches: First-order statistics and Second-order statistics. Various First-order statistics and Second-order statistics related to this work have been explained in this section.

- **Mean:** Mean describes the average value of an image.
- **Skewness:** Skewness measures asymmetry of data around the sample mean.
- **Kurtosis:** Kurtosis measures how outlier-prone a distribution is.
- **Entropy:** Entropy returns a scalar value of gray scale image.
- **Histogram:** An estimation of the probability distribution of continuous variable in an image.
- **Contrast:** Measure of intensity between pixels and its neighbors over the whole image.
- **Energy:** Energy is the sum of squared elements in the Gray Level Co-occurrence Matrix.
- **Homogeneity:** Homogeneity is the measure of closeness of the distribution of elements in the Gray Level Co-occurrence Matrix
- **Correlation:** It specifies the measure of how correlated a pixel is to its neighbor over the whole image
- **Variance:** Variance is the average of squared differences from the mean.
- **Standard Deviation:** Standard deviation is the square root of the variance.

- **First order approach:** considers only individual pixel and not a neighbor pixel and the methods are:

- **Statistical Gray level Features:** The statistical features are used to characterize the distribution of gray level in the given image. Some of the features are Min , Max, Variance, Standard deviation, Skewness, and Kurtosis [78] are given by the Equation 5.15 to 5.20 respectively:

$$\text{Min } I_{\min} = \text{Min } \{I(X, Y)\} \qquad (5.15)$$

$$\text{Max } I_{\max} = \text{Max } \{I(X, Y)\} \qquad (5.16)$$

$$\text{Variance} = \frac{1}{mn-1}\sum_{x=0}^{m-1} \sum_{y=0}^{n-1} (\mathrm{I}(\mathrm{x}, \mathrm{y}) - \mathrm{Mean})2 \qquad (5.17)$$

Where, Mean $=1/mn \sum_{x=0}^{m-1} \mathrm{I}(\mathrm{x}, \mathrm{y})$

$$\text{Standard deviation} = \sqrt{Variance} \qquad (5.18)$$

$$\text{Skewness} = 1/mn(\sigma 3)\sum_{x=0}^{m-1} \sum_{y=0}^{n-1} (\mathrm{I}(\mathrm{x}, \mathrm{y}) - \mathrm{Mean})3 \qquad (5.19)$$

$$\text{Kurtosis} = 1/mn(\sigma 4)\sum_{x=0}^{m-1} \sum_{y=0}^{n-1} (\mathrm{I}(\mathrm{x}, \mathrm{y}) - \mathrm{Mean})4 \qquad (5.20)$$

where, m and n are the number of rows and columns respectively in the given input image.

- **Histogram Features:** It describes gray level histogram moments of an image, histogram gives details about the content of the image, and the features are average intensity, average contrast, smoothness, third moment, uniformity and entropy [78]. These features are given by Equation 5.21 to 5.25 respectively:

$$\text{Average Histogram} = \frac{1}{L} \sum_{i=0}^{L-1} \mathrm{N(i)} \qquad (5.21)$$

$$\text{Smoothness} = 1 - \frac{1}{1+\sigma 2} \qquad (5.22)$$

$$\text{Third moment} = \sum_{i=0}^{L-1} \mathrm{P(Zi - Mean)3\ P(Zi)} \qquad (5.23)$$

$$\text{Uniformity} = \sum_{i=0}^{L-1} \mathrm{P(zi)} \tag{5.24}$$

$$\text{Entropy} = \sum_{\mathrm{i=0}}^{\mathrm{L\text{-}1}} \mathrm{P(Zi)} \log_2 Zi \tag{5.25}$$

- **Second Order Statistics:** Finds features based on the pixels relationship between neighboring pixels. Some of the features are:

○ **GLCM Features:** Provides information regarding the relative position of two pixels with respect to each other. It is formed by counting the number of occurrences of pixel pairs at a given distance. GLCM matrix [78] is computed using distance vector d=(x, y), and computes the probability of occurrence of gray levels in the given distance d and angle ϴ .

Formula for computing some of the GLCM features are given in Equation 5.26 to Equation 5.29:

$$\text{Contrast} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \mathrm{P(i, j)(i - j)\ 2} \tag{5.26}$$

$$\text{Energy} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \mathrm{P(i, j)\ 2} \tag{5.27}$$

$$\text{Homogeneity} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{\mathrm{p(i,j)}}{1+|\mathrm{i-j}|} \tag{5.28}$$

$$\text{Dissimilarity} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} |\mathrm{i - j}|\mathrm{P(i, j)} \tag{5.29}$$

○ **Gray Level Run Length Matrices (GLRLM) Features:** It is a two dimensional matrix in which p(i, j) is the probability of i in j elements. These features are mainly on gray level runs with various lengths. GLRLM provides details about the spatial distributions of gray level runs in the given input image with in the distance d.

The methods of both first order and second order statistics forms hybrid features and extracts unique features of an image, hence used in pattern extraction and classification.

- **Classification**

Many classification approaches, such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), k-NN (k-Nearest Neighbor), Fuzzy-based, have been widely applied for image classification. In general, image classification approaches can be grouped as Hard and Soft (fuzzy) classification, or Per-pixel, Sub-pixel, and Perfield [78].

The classifier used in the work is fuzzy based classifier which employs if-then rules for analysis. The quantitative analysis of features obtained by feature extraction module, qualitative reasoning, and human knowledge will convert the ten features into one single model value. Fuzzy models are of two types: Mamdani fuzzy model and Sugeno fuzzy model. Mamdani Fuzzy inference is the most commonly used fuzzy methodology built using fuzzy set theory. Figure 5.16 shows Fuzzy Inference System where input undergone for some rules yields inferred output, then scaled down by aggregator function. A crisp value used in last part of the classifier is extracted from Defuzzyfication. The strategies for Defuzzyfication are Centroid of the area, Bisector of the area, Mean of maximum, Smallest of maximum, and Largest of maximum.

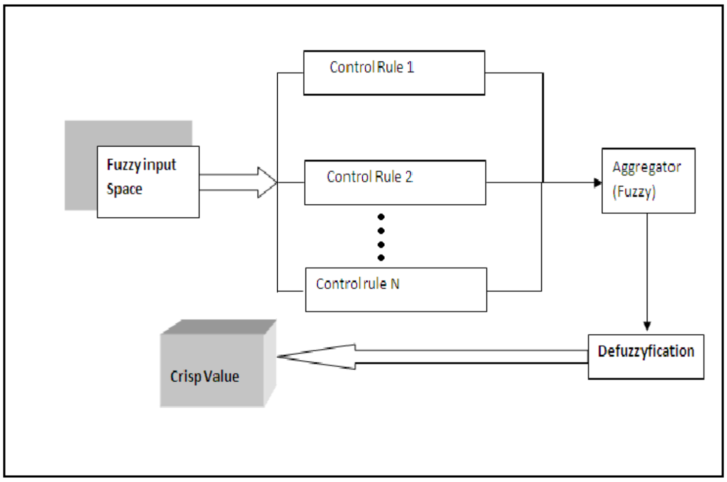


Figure 5.16: Generic Model of Fuzzy Inference System

5.10.3 Detailed Description and Algorithms

This section gives a detailed functional description and algorithm of the phases in the approach for recognition of ancient epigraphic text.

➢ Extraction of Statistical Features of the Segmented Character

The purpose of this method is to compute statistical features like one dimensional features, histogram features and GLCM features of the segmented character image. These extracted features are further used for classification and identification.

Algorithm: **Feature_Extraction (Segmented Character)**

Input

Segmented character of Brahmi or Hoysala script.

Output

The values of One-Dimensional Statistical features such as Mean, Variance, Standard deviation, the values of Histogram Features like Entropy, Skewness, Kurtosis and the values of GLCM features like Energy, Contrast, Correlation, Homogeneity are considered as output.

Method:

```
Step 1:  [Compute 1-D Statistical Features]
            Compute Mean, variance, Standard deviation
Step 2: [Compute Histogram Features]
            Compute Skewness, Kurtosis, Entropy
Step  3: [Compute GLCM features Features]
            Compute Energy,  Homogeneity, Contrast,  Correlation
```

➢ Classification and Recognition of the Segmented Character

In this work Mamdani fuzzy model is adopted which uses Gaussian based computations. Computations are adopted in two stages. First stage computation is to model and classify all the characters of Brahmi and Hoysala script which will be used as training sample to achieve the proper output. As a pre-preparation work, the first stage model value must be stored in the database. The computed model value is stored as a document in excel worksheet along with relevant modern Kannada language characters.

Second stage computation is to get the model value of features extracted after segmentation. The result of the second stage will be further used in classification and comparisons of data with the data set stored in excel worksheet.

***Algorithm*: Recognition (Character_Features)**

Purpose

The purpose of this module is to prepare training data set for future computations and to compute a model value which will be compared with the training data set to match the segmented character. Once the search is found, the relevant modern Kannada character will be showcased as output

Input : Ten features from feature extraction module

Output : Recognized Kannada character in modern form.

Method:

```
Step 1: Find the model value
Step 2: Compare the model value with classes described in dataset
Step 3: Compare model value in particular class applying if-then rules
Step 4: if (found nearest match)
            Retrieve the entire row in the dataset
            Display the result of modern Kannada character
Step 5: else
           Display “result not found”
```

Method ends.

5.10.4 Experimental Results and Analysis

This section presents the results and performance characteristics of the approach for character recognition during Ashoka and Hoysala periods.

➢ **Experimental Data**

In this work, the experimental analysis is carried out on scanned images of ancient Kannada epigraphs from Ashoka and Hoysala periods.

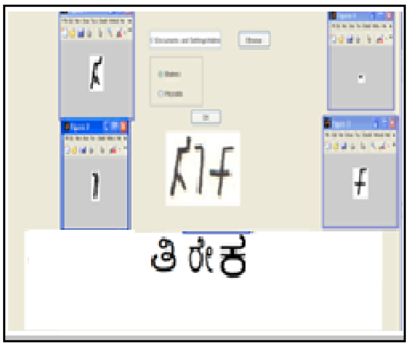
➢ **Experimental Results**

The section illustrates the results of experimentation of the current work.

Figure 5.17 shows the results of Segmentation phase for a sample input epigraphic image.

- **Results of Recognition of Brahmi Script**

Figure 5.17 shows the results of recognition when the user has selected Brahmi text image as the input. As a result, segmented characters are displayed with the recognized modern Kannada character.

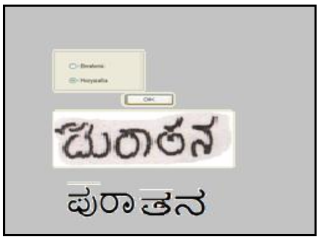


ತಿ ರೀ ಕ

Figure 5.17: Results of Recognition for Brahmi Script using Fuzzy Classifier

- **Results of Character Recognition for Hoysala Script**

Figure 5.18 shows the results of recognition for a noise-free, skewed sample input script of Hoysala period.



○ Brahmi
● Hoysala

ಪುರಾತನ

Figure 5.18: Results of Recognition for Hoysala Script using Fuzzy Classifier

- **Performance Measure**

The recognition rate of characters from Brahmi script is on an average 88% and Hoysala script is 83% .

- **Inferences from the Experimentation and Analysis**

The following inferences can be drawn from the experimentation results obtained:

- The accuracy of Recognition is comparatively high for Brahmi text, as the character set includes letters which have completely connected components.
- Hoysala text documents have many disconnected components of a single character, yielding lesser recognition rate.

5.11 Summary

This chapter discussed the methods used for recognition of epigraphical characters from different periods. Six methods for feature extraction and recognition of epigraphical characters with a combination of different classifiers are discussed.

Based on structural and statistical features with different classifiers, methods are explored for recognition of epigraphical text. The methods discussed are:- Zernike Features with SVM Classifier; Central and Zernike Moment features with RF Classifier; Zone-based and Gabor features with ANN; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier; SURF features with SVM, ANN and k-NN classifier; finally Fuzzy Classifier using First-Order and Second-order Statistical features for the for recognition of epigraphic documents. These techniques have been experimented with the test images of different periods and the results obtained are satisfactory. The performance characteristics of the approaches are also discussed.

Chapter 6

A COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS FOR RECOGNITION OF EPIGRAPHIC CHARACTERS

6.1 Premise

In Chapter 4 and Chapter 5 techniques were proposed for Classification of Epigraphic records according to period and Recognition of epigraphic characters respectively. This Chapter discusses the experimental results and a comparative study of various techniques adopted for recognition of epigraphs which includes: Central and Zernike Moment features with SVM, k-NN and RF Classifier in Section 6.2; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier in Section 6.3; and SURF Features with SVM, k-NN, ANN Classifier in Section 6.4. The chapter is summarized in Section 6.5.

6.2 Central and Zernike Moment Features with SVM, k-NN and RF Classifier

The method of dating epigraphical records and recognition using Central and Zernike Moment Features with designed Random Forest Classifier was discussed in Chapter 4 and 5 respectively. This section gives a comparative study of performance of designed RF Classifier with SVM, k-NN classifiers using Central and Zernike Moment features.

6.2.1 Evaluation Metrics

Metrics are the various measures which facilitate the quantification of some particular characteristics. The metrics used to evaluate the proposed work are:

- **Classification rate:** This metric is used to determine the accuracy of the Classifier, which is given by the number of correct character classifications out of the total number of character samples in the document.

$$Classification\ Rate = \frac{Number\ of\ Correctly\ Classified\ Characters}{Total\ Number\ of\ Characters\ in\ input\ document}$$

- **Training time**: This metric measures the time taken to train the Classifier.
- **Classification time:** The classification time is the time taken to predict the class labels for the given set of inputs.

6.2.2 Performance Analysis of RF with SVM and k-NN

A comparison of classification rates of designed RF Classifier with SVM and k-NN using Central and Zernike moments is made on epigraphic base characters from Ashoka, Satavahana, and Kadamba periods.

The designed RF Classifier is tested on 105 handwritten samples with 35 base characters, belonging to each of the dynasties:- Ashoka, Kadamba, and Satavahana dynasty.

These handwritten characters are also tested with the SVM and k-NN classifiers provided by the OpenCV library. The parameters used for the classifiers are as follows:

SVM classifier: Kernel type – Linear; SVM type – C-SVM (type 1)

k-NN classifier: k = 5

RF classifier: Number of trees = 30; number of thresholds = 10

- **Classification of Characters from Ashoka period**

The classification rates for ancient Kannada characters from Ashoka period are tabulated in Table 6.1. The number of samples taken for training is 70 and 35 for testing. The accuracy in the classification of characters is compared using the designed RF with SVM and k-NN.

Figure 6.1 shows the corresponding plot representing the classification rates using RF, SVM and k-NN, for characters from Ashoka period.

Table 6.1: Classification Rates of Characters from Ashoka Period using SVM, k-NN and RF

Classifier		**Training (%)**	**Testing (%)**
	SVM	88.41	77.13
	k-NN	56.52	44.44
	RF	94.20	69.22

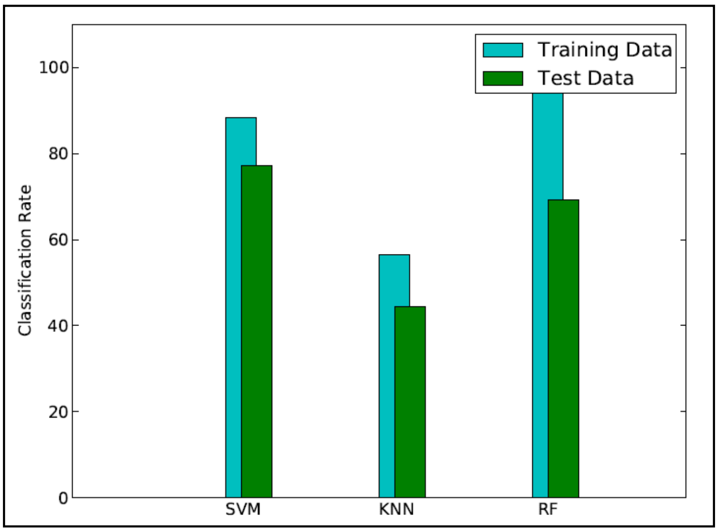


Figure 6.1: The Classification Accuracy of Characters from Ashoka Period using SVM, k-NN and RF

➢ Classification of Characters from Kadamba period

The classification accuracy- of the designed RF is compared with SVM and k-NN, for handwritten ancient Kannada characters from Kadamba period and is tabulated as in Table 6.2. The number of samples taken for training is 70 and 35 for testing. Figure 6.2 shows the corresponding plot representing the classification rates using RF, SVM and k-NN, for characters from Kadamba period.

Table 6.2: Classification Rates of Characters from Kadamba Period using SVM, k-NN and RF

		Training (%)	Testing (%)
Classifier	**SVM**	77.62	69.94
	k-NN	52.23	48.18
	RF	87.15	65.12

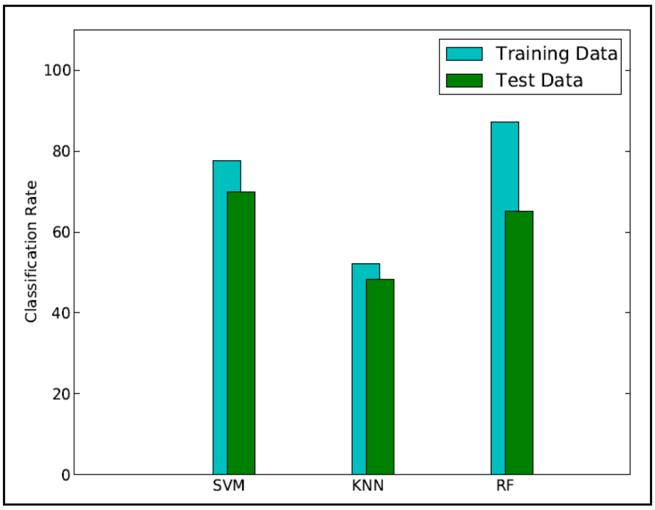


Figure 6.2: The Classification Rates of Characters from Kadamba Period using SVM, k-NN and RF

- **Classification of characters from Satavahana period**

The classification rates for handwritten ancient Kannada characters from Satavahana period are tabulated in Table 6.3. The number of samples taken for training is 70 and 35 samples is taken for testing. The accuracy in the classification of characters using the designed RF with SVM and k-NN is being compared.

Figure 6.3 represents the corresponding plot of accuracy, in classification of characters from Satavahana period, using RF, SVM and k-NN

Table 6.3: Classification Rates of Characters from Satavahana Period using SVM, k-NN and RF

Classifier		Training (%)	Testing (%)
	SVM	81.54	74.12
	k-NN	58.46	52.94
	RF	90.77	67.06

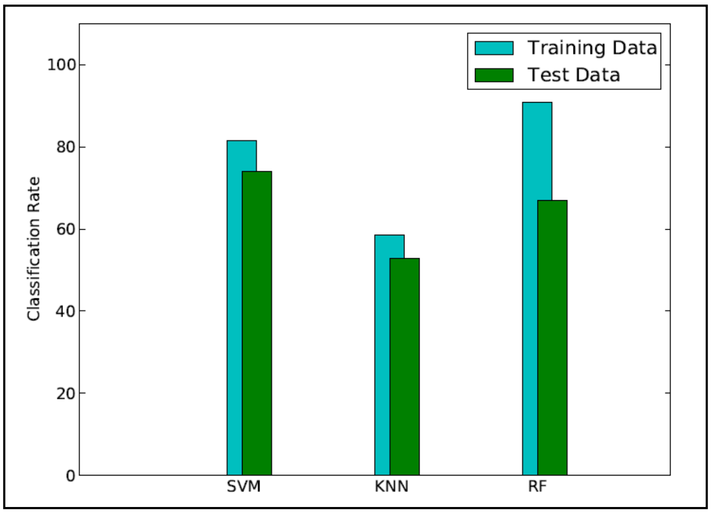


Figure 6.3: Classification Rates of Characters from Satavahana Period using SVM, k-NN and RF

6.2.3 Inference from the Performance Analysis

The following inferences are drawn from the analysis of performance of Classifiers:

- RF Classifier works better in classifying the trained characters. Hence, a well trained RF classifier will perform comparatively better than SVM and k-NN in the classification of new characters.
- SVM classifier works well in the classification of new characters. Hence, SVM performs comparatively better than RF and k-NN- when the classifier is not trained with sufficient samples.
- K-NN Classifier’s capability lies in between RF and SVM in the classification of trained and test data.

6.3 Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier

The method of classification of epigraphical characters using Fourier features was discussed in the previous chapter. This section gives a comparative study of the performance of SVM, k-NN, ANN and Naive Bayes Classifiers using Fourier Features.

6.3.1 Evaluation Metrics

Metrics are the various measures which facilitate the quantification of some particular characteristics provided by Confusion Matrix shown in Table 6.4.

Table 6.4: Generic Confusion Matrix for Classification

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

The metrics used to evaluate the approach are:

- **Classification rate:** This metric is used to determine the accuracy of Classifier and is given by Equation 6.1:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6.1}$$

- **Recall:** This metric is the proportion of positive cases that are correctly identified, and is calculated using Equation 6.2:

$$recall = \frac{TP}{TP+FN} \tag{6.2}$$

- **Precision:** This metric is the proportion of the predicted positive cases that are correct, and is computed using the Equation 6.3:

$$precision = \frac{TP}{TP + FP} \qquad (6.3)$$

- **Specificity:** This metric measures the proportion of negatives which are correctly identified, and is determined using the Equation 6.4:

$$specificit\,y = \frac{TN}{TN + FP} \qquad (6.4)$$

- **Training time:** This metric measures the time taken to train the Classifier.
- **Classification time:** The classification time is the time taken to predict the class labels for the given set of inputs.

6.3.2 Experimental Dataset

The character dataset of Ashoka and Hoysala period are considered for recognition of ancient Kannada epigraphs using Fourier features. The system is trained with 258 characters of Ashoka period and 362 characters of Hoysala period with 4 different instances of each. The system is tested with the 40 reconstructed epigraphs containing Ashoka and Hoysala characters.

6.3.3 Performance Analysis

The performance analysis is made for the 40 epigraphical images containing 250 Ashoka and 200 Hoysala characters. The performance characteristics of SVM, ANN, k-NN, NB Classifiers using Fourier features are graphed in Figure 6.4.

The Confusion Matrix for the SVM, ANN, k-NN and NB classification are shown in Table 6.5(a) – 6.5(d) respectively. The indicated values in the tabulation represent the number of images.

The various Performance Metrics (Accuracy, Precision, Recall, and Specificity) for the test data are shown in Table 6.6. The tabulations are shown in percentage, each column indicates the Classifier used and the rows indicate the Metric value.

Table 6.5: Confusion Matrix for SVM, ANN, k-NN and NB Classification

(a) Confusion Matrix for SVM Classification

		Ashoka		**Hoysala**	
		Positive	**Negative**	**Positive**	**Negative**
Predicted	**Positive**	182	18	138	22
	Negative	23	27	17	23

(b) Confusion Matrix for ANN Classification

		Ashoka		**Hoysala**	
		Positive	**Negative**	**Positive**	**Negative**
Predicted	**Positive**	166	34	125	35
	Negative	24	26	22	18

(c) Confusion Matrix for k-NN Classification

		Ashoka		**Hoysala**	
		Positive	**Negative**	**Positive**	**Negative**
Predicted	**Positive**	104	96	88	72
	Negative	31	19	31	9

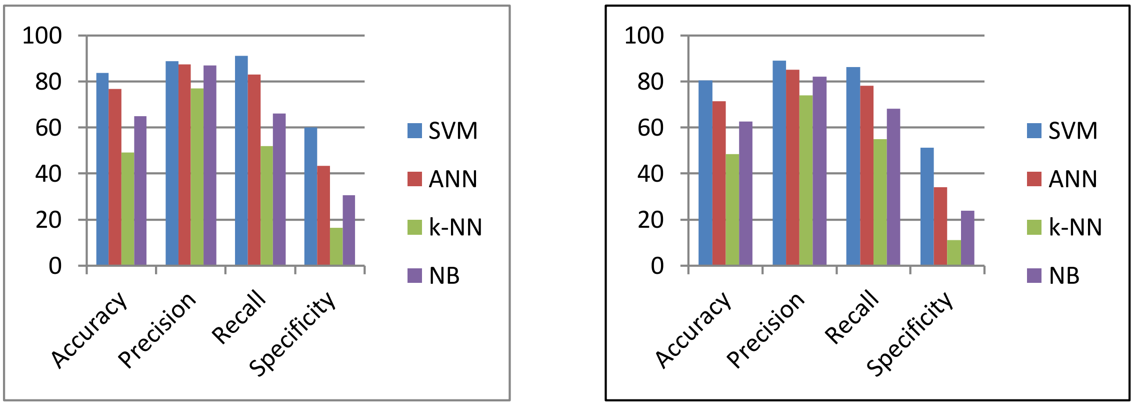
(d) Confusion Matrix for NB Classification

		Ashoka		**Hoysala**	
		Positive	**Negative**	**Positive**	**Negative**
Predicted	**Positive**	132	68	109	51
	Negative	20	30	24	16

Table 6.6: Performance Metrics (in percentage) for Test data using SVM, ANN, k-NN and NB Classifier

		Ashoka				Hoysala			
		SVM	**ANN**	**k-NN**	**NB**	**SVM**	**ANN**	**k-NN**	**NB**
Metrics	**Accuracy**	83.60	76.80	49.20	64.80	80.50	71.50	48.50	62.50
	Precision	88.78	87.36	77.03	86.84	89.03	85.03	73.94	81.95
	Recall	91.00	83.00	52.00	66.00	86.25	78.12	55.00	68.12
	Specificity	60.00	43.33	16.52	30.61	51.11	33.96	11.11	23.88

The graph of the Table 6.6 is shown in Figure 6.4 (a) and Figure 6.4 (b).



(a) Performance metrics of Classifiers (Ashoka)

(b) Performance metrics of Classifiers (Hoysala)

Figure 6.4: Performance Metrics (in percentage) for Test data using SVM, ANN, k-NN and NB Classifier

6.3.4 Inference from the Performance Analysis

The following inferences are drawn based on the performance analysis of classifiers:

- All the classifiers show better performance for character recognition of Ashoka period compared to Hoysala period.

- SVM Classifier works better in classifying the test data. ANN Classifier's capability lies in between those SVM and NB Classifiers, k-NN have lowest classification rate.

6.4 SURF Features with SVM, k-NN and ANN Classifier

In Chapter 5, OCR system was developed to recognize the ancient scriptures of Ashoka period using SURF features. This section covers the performance evaluation of the system based on the classification accuracy of ancient records.

6.4.1 Experimental Dataset

The dataset on which the experiment was conducted are the character samples from Ashoka period. The character set of Ashoka period consists of 7 vowels, 33 consonants, and 218 compound characters. The Test dataset for classification includes:- 5 samples of 20 randomly chosen Ancient Kannada Characters that are segmented from the inputted document.

6.4.2 Performance Analysis

The performance of SVM, k-NN, ANN and Multiple classifiers with SURF features for recognition of ancient records is elucidated in this section. The recognition accuracy of the classifiers with varying data sizes is represented graphically. The performance of the three classifiers individually, as well as when they are combined is compared and illustrated pictorially.

SVM's behavior is analyzed with varying dataset size and its accuracy is noted and is depicted in Figure 6.5. The average accuracy of ANN under different dataset size is 92.16%.

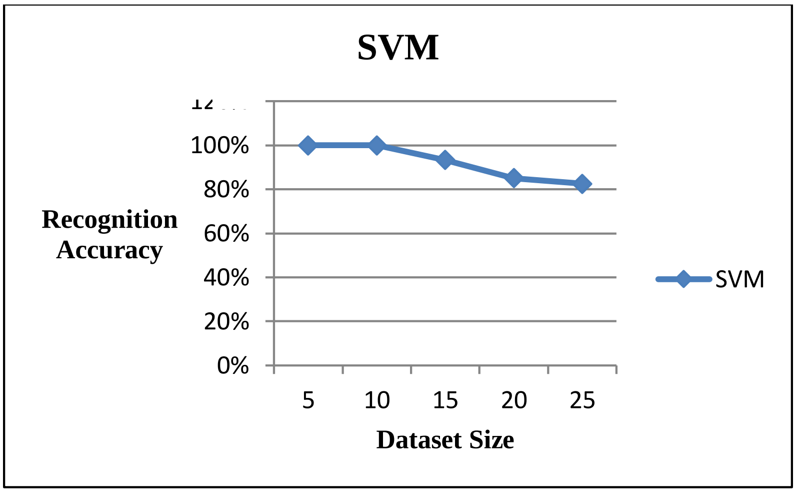


Figure 6.5: SVM Recognition Accuracy using SURF features

The behavior of k-NN is shown in Figure 6.6. The average accuracy of k-NN under different dataset size is 91.66%.

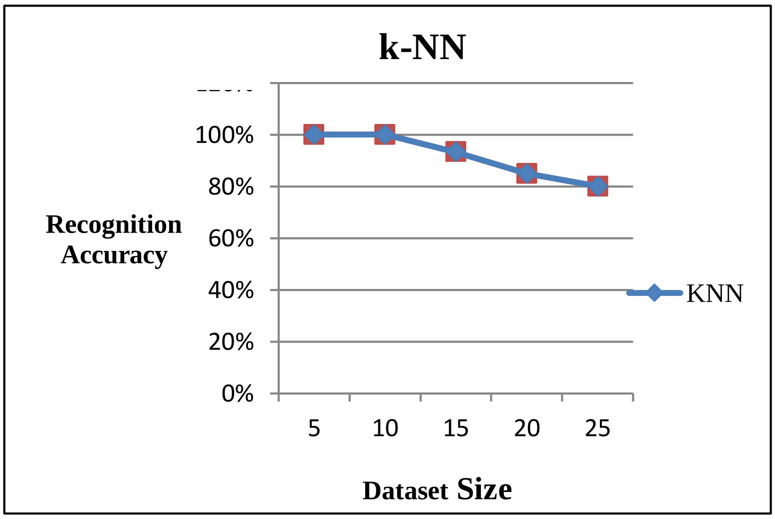


Figure 6.6: k-NN Recognition Accuracy using SURF features

ANN’s behavior is analyzed with varying dataset size and its accuracy is noted and is depicted in Figure 6.7. The average accuracy of ANN under different dataset size is 92.6%.

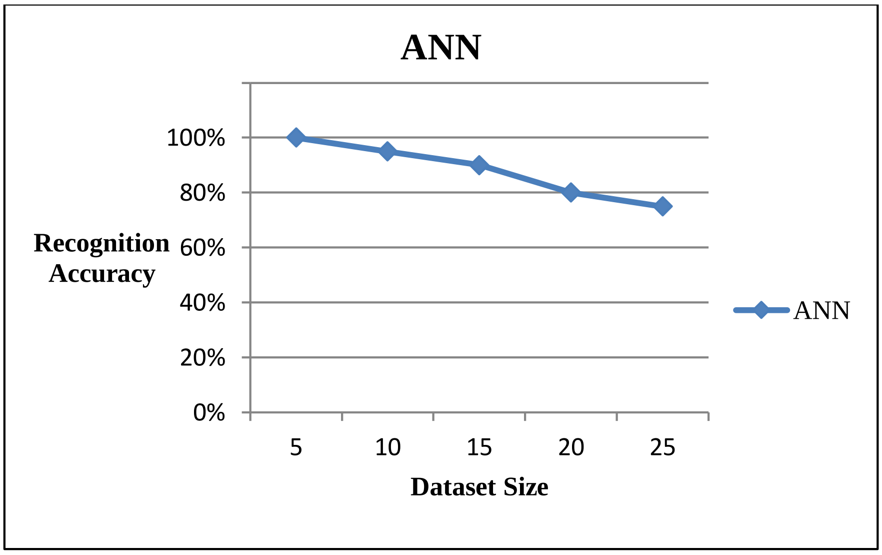


Figure 6.7: ANN Recognition Accuracy using SURF features

The output of three Classifiers SVM, k-NN and ANN is passed to the Confusion Matrix. The Confusion Matrix gives the recognized character. The output from the Confusion Matrix is analyzed and the results of recognition accuracy are indicated in the graph as shown in Figure 6.8. The average accuracy of combining the classifiers is 94.66%.

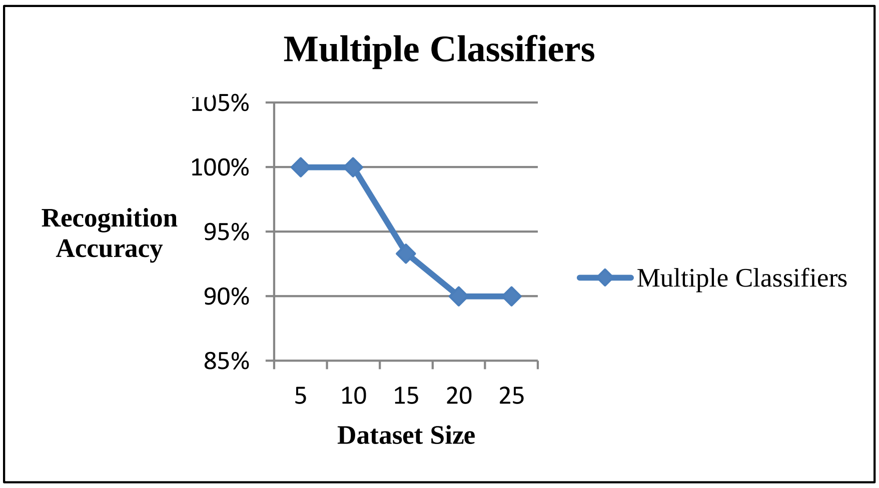


Figure 6.8: Multiple Classifiers Recognition Accuracy

The bar graph shown in Figure 6.9 gives the comparison of all classifiers SVM, k-NN and ANN for different dataset sizes. It is observed that with the combination of classifiers the recognition rate increases considerably.

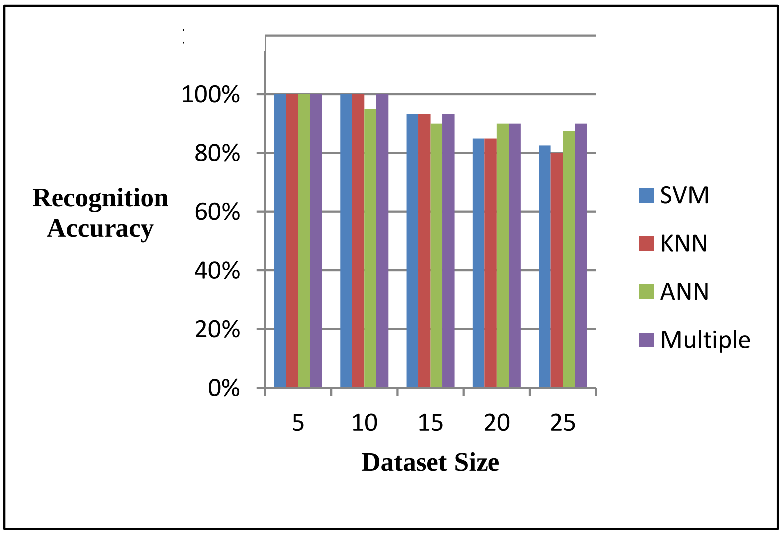


Figure 6.9: Comparison of Recognition Accuracy of SVM, k-NN, ANN and Multiple classifiers for different dataset size

6.4.3 Inference from the Performance Analysis

Following inferences can be drawn from the Experimental Analysis:

- SVM Classifier works better for small training set compared to ANN and k-NN.
- ANN, when compared to SVM and k-NN, performs better when the size of the dataset increases.
- The performance of k-NN lies between SVM and ANN.
- For any dataset size, the accuracy of recognition of ancient Kannada characters improves when multiple classifiers are used as compared to a single classifier recognizing the characters.

6.5 Summary

This chapter covered a comparative experimental study of various techniques adopted for recognition of epigraphic characters. At first, a comparative study on the performance of designed RF with SVM and k-NN using Central and Zernike moment features is carried out, in the classification of ancient characters from Ashoka, Satavahana and Kadamba dynasties. Secondly, the performance characteristics of SVM, ANN, k-NN and NB Classifiers using Fourier features for the classification of ancient characters from Ashoka, and Hoysala periods is observed. The strengths and weaknesses of these classification methods are also discussed. Thirdly, the behavior of SVM, k-NN, ANN and Multiple classifiers using SURF features is analyzed. It is observed that the recognition accuracy increases when multiple classifiers are used along with the confusion matrix, which resolves the errors which may arise during character recognition.

Chapter 7

CONCLUSION AND FUTURE RESEARCH AVENUES

7.1 Summary

Epigraphic studies are of great significance to mankind. Historical records are degraded or damaged over-time and hence preservation of this is important. Manual efforts of expert epigraphers in reading the inscriptions are time consuming and confronted with many problems. Also, common laymen find difficulty in reading these ancient epigraphs. Hence, in the current research work efforts are made for automatic recognition of epigraphic documents.

The research work focuses on the development of a character recognition system for ancient epigraphical documents. In this thesis, as presented in previous chapters the automatic epigraphic character recognition system is designed and implemented in three phases: *Preprocessing and Segmentation*, *Classification of epigraphic scripts* according to period and *recognition of epigraphic characters*.

An introduction to the research work with a brief overview of the domain, epigraphy and its relevance, related work, motivation, and objectives is presented in Chapter 1.

In Chapter 2 the state-of-the-art development in the field of epigraphical document analysis and recognition is discussed. An overview of Indian epigraphy, the evolution of Indian scripts, epigraphical documents - its sources and challenges encountered in decipherment is covered.

Chapter 3 on Preprocessing and Segmentation of Epigraphs explores existing techniques of Spatial Filtering and Noise Elimination for enhancing epigraphical documents, followed by binarization. It also presents the algorithmic models for segmentation of epigraphical documents. The experimental results and performance analysis are demonstrated.

Chapter 4 on Classification of ancient Kannada Epigraphs into different periods presents two models for dating ancient epigraphical records: SVM Classifier based method using Zonal Features and the design of Random Forest Classifier with Central and Zernike moment features. The experimental results and performance analysis of these models are illustrated.

In Chapter 5 approaches for recognition of epigraphical characters with different models to decipher text of ancient times in a combination of varying feature extraction techniques are explored. The experimental results and performance analysis of these approaches for recognition of epigraphical documents are presented.

In Chapter 6, a comparative study and performance analysis of different classifiers for recognition of epigraphical records with varying feature extraction techniques is made.

Lastly, Chapter 7 provides the concluding remarks highlighting the major contributions towards this research work and future research avenues in automation of deciphering epigraphical documents.

7.2 Significant Contribution of this thesis

The key highlights of this work are enumerated as follows:

- Study and analysis of different available techniques on preprocessing ancient epigraphic documents which include enhancement using spatial filtering, noise elimination and binarization.

- Explore different existing approaches for segmentation of epigraphical documents

- Determination of period of input epigraph and performance analysis using two approaches: SVM Classifier based method with Zonal Features; and design of Random Forest Classifier for predicting a period of ancient Kannada epigraphs using Normalized Central Moments and Zernike Moments features.

- Automatic recognition of epigraphic records and performance analysis with different combination of feature extraction techniques and classifiers such as :- Zernike Features with SVM Classifier; Central and Zernike Moment features with RF Classifier; Zone-based and Gabor features with Neural Network approach; Fourier Features with SVM, k-NN, ANN and Naive Bayes Classifier; SURF features with SVM, ANN, and k-NN classifiers; and finally Fuzzy Classifier is designed and implemented for recognition of ancient scripts using First-order and Second-order Statistical features.

- A Comparative study and performance analysis of different classifiers with varying feature extraction techniques for recognition of epigraphs such as :- Central and Zernike Moment features with SVM, k-NN and RF Classifier; Fourier Features with S V M, k- NN, ANN and Naive Bayes Classifier; and SURF Features with SVM, k-NN, ANN and Multiple Classifier.

The current work finds scope in the department of Epigraphy, Ancient History and Archaeology for automated reading of ancient scripts. Thus, the work assists epigraphers, archaeologists and historians for digitization and further exploration of ancient historical records.

7.3 Future Research Avenues

In this research work efforts are made to classify and recognize ancient Kannada epigraphs of different times. The design of an automated epigraphical document recognition system with reasonable accuracy- is imperative.

Some of the challenges in automatic reading of epigraphs which can be addressed further are listed below:

- To strengthen the techniques of preprocessing so as to handle raw epigraphic images of varying qualities from different sources, so that a noise-free epigraph can be input for classification and recognition phases.

- Devise better segmentation methods to address the structural complication, non-uniform spacing, skew, touching lines as well as characters, erased letters, broken characters, overlapping characters.

- To extend the proposed models for classification and recognition of epigraphic records to any Indian or non-Indian epigraphical scripts, and achieve better performance results.

- Creating standard / bench mark data corpus of ancient characters of a script from different periods to achieve better accuracy for classification and recognition.

- Parallel classification and recognition of epigraphic documents to achieve the speedup in the process, once desirable recognition rate is attained.

Publications in Support of this Thesis

Journals

1. "Recognition of Historical Documents using Gabor and Zonal Features*",* In *Signal & Image Processing: An International Journal (SIPIJ),* vol.6, no.4, pp. 57-69, 2015.

2. "Performance Analysis of Random Forests with SVM and k-NN in Classification of Ancient Kannada Scripts", In *International Journal Of Computers and Technology(IJCT),* vol. 13, no. 9, pp. 4907-4921, 2014.

3. "Preprocessing of Camera Captured Inscriptions and Segmentation of Handwritten Kannada text*",* In *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE),* vol. 3, no. 5, pp. 6794-6803, 2014.

4. "Zernike Moment features for the Recognition of Ancient Kannada Base Characters", In *International Journal of Graphics & Image Processing (IJGIP),* vol. 4, no. 2, pp. 99-104, 2014.

5. "SVM Classifier for the Prediction of Era of an Epigraphical Script", In the *International Journal of Peer to Peer Networks (IJP2P)* vol.2, no.2, pp. 12-22, 2011.

6. "Automatic Decipherment of Ancient Indian Epigraphical Scripts - A Brief Review", In *International Journal of Computer Science & Emerging Technologies* (E-ISSN: 2044-6004), vol. 2, no. 1, pp. 139-143, Feb 2011.

Conference Proceedings

7. "Fourier Features for the Recognition of Ancient Kannada Text", In the *Proceedings of Second International Conference on Computational Intelligence in Data Mining (ICCIDM-2015)* – vol. 1, Advances in Intelligent Systems and Computing 410, *Springer,* Odisha, pp. 421-428, 2015.

8. "Enhancement and Segmentation of Historical Records", In the *Proceedings of Fifth International Conference on Digital Image Processing and Pattern*

Recognition (DPPR 2015), Computer Science Conference Proceedings – AIRCC publications, Chennai, vol. 5, no.13, pp. 95-113, 2015.

9. "Feature Extraction and Recognition of Ancient Kannada Epigraphs", In the *Proceedings of International Conference on Computational Intelligence in Data Mining (ICCIDM-2014)*- Smart Innovation, Systems and Technologies 33, *Springer*, Veer Surendra Sai University of Technology(VSSUT), Burla, Odisha, vol.3, pp. 469-477, 2014.

10. "Recognition of Ancient Kannada Epigraphs using Fuzzy-Based Approach", In the *Proceedings of International Conference on Contemporary Computing and Informatics (IC3I-2014), IEEE*, SJCE, Mysore, pp. 657-662, 2014.

11. "Classification of ancient epigraphs into different periods using Random Forests*",* In the *Proceedings of Fifth International Conference on Signal and Image Processing (ICSIP-2014), IEEE*, BNMIT, Bangalore, pp.172-178, 2014.

12. "Dating of Ancient Epigraphs using Random Forest Classifier", In the *Proceedings of International Conference on Emerging Computation and Information Technologies (ICECIT-2013), Elsevier*, Siddaganga Institute of Technology, Tumkur, pp. 331-339, 2013.

Bibliography

1. A.V.Narasimha Murthy, “Kannada Lipiya Ugama Mattu Vikasa’, Published by Kannada Adhyayana Samsthe, Mysore University, Mysore, 1968.
2. M G Manjunatha, G K Devarajaswamy, “ Kannada Lipi Vikasa”, 1st edition, Published by Mantralaya, Jagadguru Sri Manmadhwacharya Mula Mahasamstana, Sri Ragavendra Swamy Mutt, 2004.
3. Dr. Devarakonda Reddy, “ Lipiya Huttu Mattu Belavanige”- Origin and Evolution of Script, Published by Kannada Pustaka Pradhikara (KannadaBook Authority), Bangalore.
4. D Dayalan - Computer Application in Indian Epigraphy, Bharatiya Kala Prakashan publication, 2005.
5. Srihari, Sargur N., Yong-Chul Shin, Vemulapati Ramanaprasad, and Zhixin Shi. "Document image-processing system for name and address recognition." *International journal of imaging systems and technology* 7, no. 4 (1996): 379-391.
6. Dori, Dov, David Doermann, Christian Shin, Robert Haralick, Ihsin Phillips, Mitchell Buchman, and David Ross. "The representation of document structure: A generic object-process analysis." *Handbook of character recognition and document image analysis* (1997): 421-456.
7. Sastry, Panyam Narahari, and Ramakrishnan Krishnan. "Isolated Telugu Palm leaf character recognition using Radon Transform—A novel approach." In *Information and Communication Technologies (WICT), 2012 World Congress on*, pp. 795-802. IEEE, 2012.
8. Boussellaa, Wafa, Aymen Bougacha, Abderrazak Zahour, Haikal El Abed, and Adel Alimi. "Enhanced text extraction from Arabic degraded document images using EM algorithm." In *2009 10th International Conference on Document Analysis and Recognition*, pp. 743-747. IEEE, 2009.
9. Sreedevi, Indu, Rishi Pandey, Geetanjali Bhola, and Santanu Chaudhury. "Enhancement of inscription images." In *Communications (NCC), 2013 National Conference on*, pp. 1-5. IEEE, 2013.
10. Sreedevi, Indu, Rishi Pandey, N. Jayanthi, Geetanjali Bhola, and Santanu Chaudhury. "NGFICA based digitization of historic inscription images." *ISRN Signal Processing* 2013 (2013).

11. Karthik, S., H. R. Mamatha, and Srikanta Murthy. "An Approach based on Run Length Count for Denoising the Kannada Characters." *International Journal of Computer Applications* 50, no. 18 (2012).
12. Gangamma, B., K. Srikanta Murthy, and Arun Vikas Singh. "Restoration of degraded historical document image." *Journal of Emerging Trends in Computing and Information Sciences* 3, no. 5 (2012): 36-39.
13. RajaKumar, S., and V. Subbiah Bharathi. "Ancient Tamil Script Recognition from Stone Inscriptions Using Slant Removal Method." In *IEEE International Conference on Electrical, Electronics and Biomedical Engineering* (2012)
14. Su, Bolan, Shijian Lu, and Chew Lim Tan. "Binarization of historical document images using the local maximum and minimum." In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 159-166. ACM, 2010
15. Messaoud, Ines Ben, Hamid Amiri, Haikal El Abed, and Volker Märgner. "Region Based Local Binarization Approach for Handwritten Ancient Documents." In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pp. 633-638. IEEE, 2012.
16. Devi, H. "Thinning: A Preprocessing Technique for an OCR System for the Brahmi Script." *Ancient Asia* 1 (2006).
17. K. Srikanta Murthy, G.Hemantha Kumar, P.Shivakumar, "A Novel method based on rectangle fitting for noise removal in an epigraphical script," *Proceedings of 39th Annual convention of Computer Society of India*, Mumbai, pp 166-171, 2004.
18. Doreswamy, K.Srikanta Murthy, G.Hemantha Kumar, P.Nagabhushan "Filtering Technique based on Minimum Majority Function to eliminate noises from the Epigraphical Script Images " *Proceedings of National Conference on Recent Trends in Information Technology*, Tamil Nadu, pp.no 35-39, 2002.
19. Van Phan, Truyen, Bilan Zhu, and Masaki Nakagawa. "Collecting handwritten Nom character patterns from historical document pages." In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pp. 344-348. IEEE, 2012.

20. Chamchong, Rapeeporn, and Chun Che Fung. "Text line extraction using adaptive partial projection for palm leaf manuscripts from Thailand." In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pp. 588-593. IEEE, 2012.
21. Rao, Adabala Venkata Srinivasa. "Segmentation of Ancient Telugu text documents." *International Journal of Image, Graphics and Signal Processing* 4, no. 6, pp 8-14, 2012.
22. K. Srikanta Murthy, G.Hemantha Kumar, P.Shivakumar, P.R.Ranganath, "Nearest Neighbor clustering based approach for line and character segmentation in epigraphical scripts", Proceedings of International conference on cognitive systems (ICCS-2004), New Delhi, pp. 1-5, 2004.
23. Doreswamy, K.Srikanta Murthy, G.Hemantha Kumar, P.Nagabhushan "Partial Eight Direction Based Algorithm for Line Segmentation of Epigraphical Script Images" Proceedings of National Conference ETA-2003, Saurashtra University, Rajkot, July 11-13, 2003.
24. Ramappa, Mamatha Hosalli, and Srikantamurthy Krishnamurthy. "Skew Detection, Correction and Segmentation of Handwritten Kannada Document."*International Journal of Advanced Science and Technology* 48 (2012).
25. Kang, Le, David Doermann, Huaigu Cao, Rohit Prasad, and Prem Natarajan. "Local segmentation of touching characters using contour based shape decomposition." In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pp. 460-464. IEEE, 2012.
26. Kumar, S. Raja, and V. Subbiah Bharathi. "An off line ancient tamil script recognition from temple wall inscription using Fourier and Wavelet features."*Eur. J. Sci. Res* 80, no. 4 (2012): 457-464.
27. Alirezaee, Shahpour, Hassan Aghaeinia, Majid Ahmadi, and Karim Faez. "Recognition of middle age Persian characters using a set of invariant moments." In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, pp. 196-201. IEEE, 2004.
28. Sridevi, N., and P. Subashini. "Combining Zernike moments with regional features for classification of handwritten ancient Tamil scripts using Extreme learning machine." In *Emerging Trends in Computing, Communication and*

Nanotechnology (ICE-CCN), 2013 International Conference on, pp. 158-162. IEEE, 2013.

29. Bandara, Dammi, Nalin Warnajith, Atsushi Minato, and Satoru Ozawal. "Creation of precise alphabet fonts of early Brahmi script from photographic data of ancient Sri Lankan inscriptions." *Can. J. Artif. Intell. Mach. Learn. Pattern Recognit* 3, no. 3 (2012): 33-39.
30. Meza-Lovon, Graciela Lecireth. "A graph-based approach for transcribing ancient documents." In *Ibero-American Conference on Artificial Intelligence*, pp. 210-220. Springer Berlin Heidelberg, 2012.
31. Zahedi, M., and S. Eslami. "Improvement of Random Forest Classifier through Localization of Persian Handwritten OCR." *ACEEE Int. J. Inf. Technol* 1, no. 2 (2012): 31-36.
32. Azmi, Mohd Sanusi, Khairuddin Omar, Mohammad Faidzul Nasrudin, Azah Kamilah Muda, and Azizi Abdullah. "Digital paleography: Using the digital representation of Jawi manuscripts to support paleographic analysis." In *Pattern Analysis and Intelligent Robotics (ICPAIR), 2011 International Conference on*, vol. 1, pp. 71-77. IEEE, 2011.
33. Wolf, Lior, Liza Potikha, Nachum Dershowitz, Roni Shweka, and Yaacov Choueka. "Computerized paleography: tools for historical manuscripts." In *2011 18th IEEE International Conference on Image Processing*, pp. 3545-3548. IEEE, 2011.
34. Zaghden, Nizar, Remy Mullot, and Adel M. Alimi. "Characterization of ancient document images composed by Arabic and Latin scripts." In *Innovations in Information Technology (IIT), 2011 International Conference on*, pp. 124-127. IEEE, 2011.
35. Papaodysseus, Constantin, Panayiotis Rousopoulos, Dimitris Arabadjis, Fivi Panopoulou, and Michalis Panagopoulos. "Handwriting automatic classification: application to ancient Greek inscriptions." In *Autonomous and Intelligent Systems (AIS), 2010 International Conference on*, pp. 1-6. IEEE, 2010.
36. Rashid, Sheikh Faisal, Faisal Shafait, and Thomas M. Breuel. "Connected component level multiscript identification from ancient document images." In

Proceedings of the 9th IAPR Workshop on Document Analysis System, pp. 1-4. 2010.

37. Garz, Angelika, and Robert Sablatnig. "Multi-scale texture-based text recognition in ancient manuscripts." In *Virtual Systems and Multimedia (VSMM), 2010 16th International Conference on*, pp. 336-339. IEEE, 2010.
38. Ahmad, Riaz, Syed Hassan Amin, and Mohammad AU Khan. "Scale and rotation invariant recognition of cursive Pashto script using SIFT features." In *Emerging Technologies (ICET), 2010 6th International Conference on*, pp. 299-303. IEEE, 2010.
39. Garz, Angelika, Markus Diem, and Robert Sablatnig. "Detecting text areas and decorative elements in ancient manuscripts." In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pp. 176-181. IEEE, 2010.
40. Gilliam, Tara, Richard C. Wilson, and John A. Clark. "Scribe identification in medieval English manuscripts." In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 1880-1883. IEEE, 2010.`
41. Yadav, Nisha, Hrishikesh Joglekar, Rajesh PN Rao, Mayank N. Vahia, Ronojoy Adhikari, and Iravatham Mahadevan. "Statistical analysis of the Indus script using n-grams." *PLoS One* 5, no. 3 (2010): e9506.
42. Siddiqi, Imran, Florence Cloppet, and Nicole Vincent. "Contour based features for the classification of ancient manuscripts." In *Conference of the International Graphonomics Society*, pp. 226-229. 2009.
43. Bernard, Simon, Sebastien Adam, and Laurent Heutte. "Using random forests for handwritten digit recognition." In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, pp. 1043-1047. IEEE, 2007.
44. Sousa, J. M. C., Joao Rogerio Caldas Pinto, Claudia S. Ribeiro, and Joao M. Gil. "Ancient document recognition using fuzzy methods." In *The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05.* 2005.
45. Gatos, Basilios, Kostas Ntzios, Ioannis Pratikakis, Sergios Petridis, Thomas Konidaris, and Stavros J. Perantonis. "An efficient segmentation-free approach to assist old Greek handwritten manuscript OCR." *Pattern analysis and applications* 8, no. 4 (2006): 305-320.

46. Kashyap, K. Harish, and P. A. Koushik. "Hybrid neural network architecture for age identification of ancient Kannada scripts." In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on*, vol. 5, pp. V-661. IEEE, 2003.

47. Sastry, Panyam Narahari, Ramakrishnan Krishnan, and Bhagavatula Venkata Sanker Ram. "Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach." *J. Applied Engn. Sci* 5, no. 3 (2010): 22-32.

48. Andrew, C. "Building decision trees with the ID3 algorithm." *Dr. Dobbs Journal* (1996).

49. Details of Epigraphy and its Relevance, Kannada Epigraphy :
http://www. karnatakaitihasaacademy.org

50. Details of evolution of Kannada Script and language:
http:// www.classicalkannada.org

51. Epigraphy Documents :Archaeological Survey of India (ASI) : http://asi.nic.in

52. Epigraphy Documents : Indian Council of Historical Records (ICHR) http://ichr.ac.in

53. H R, Mamatha, Sonali Madireddi, and Srikanta Murthy K. "Performance analysis of various filters for De-noising of Handwritten Kannada documents." *International Journal of Computer Applications* 48, no. 12 (2012): 30-38.

54. Gupta, Gajanand. "Algorithm for image processing using improved median filter and comparison of mean, median and improved median filter." *International Journal of Soft Computing and Engineering (IJSCE)* 1, no. 5 (2011): 304-311.

55. Trentacoste, Matthew, Ratal Mantiuk, Wolfgang Heidrich, and Florian Dufrot. "Unsharp masking, countershading and halos: Enhancements or artifacts" In *Computer Graphics Forum*, vol. 31, no. 2pt3, pp. 555-564. Blackwell Publishing Ltd, 2012.

56. Al-Ameen, Zohair, Ghazali Sulong, and Md Gapar Md Johar. "A comprehensive study on fast image deblurring techniques." *International Journal of Advanced Science and Technology* 44 (2012).

57. Yu, Zeyun, and Chandrajit Bajaj. "A fast and adaptive method for image contrast enhancement." In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 2, pp. 1001-1004. IEEE, 2004.
58. Noboyuki, Otsu. "A Threshold Selection Method from gray Level histogram."*IEEE Transactions on Systems, Man and Cybernetics* 9 (1979): 62-66.
59. Gupta, Maya R., Nathaniel P. Jacobson, and Eric K. Garcia. "OCR binarization and image pre-processing for searching historical documents." *Pattern Recognition* 40, no. 2 (2007): 389-397.
60. Bernsen, John. "Dynamic thresholding of grey-level images." In *International conference on pattern recognition*, vol. 2, pp. 1251-1255. 1986.
61. W. Niblack, An Introduction to Digital Image Processing. Englewood Cliffs,NJ:Prentice-Hall, 1986.
62. Sauvola, Jaakko, and Matti Pietikainen. "Adaptive document image binarization." *Pattern recognition* 33, no. 2 (2000): 225-236.
63. Savakis, Andreas E. "Adaptive document image thresholding using foreground and background clustering." In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, pp. 785-789. IEEE, 1998.
64. Khurshid, Khurram, Imran Siddiqi, Claudie Faure, and Nicole Vincent. "Comparison of Niblack inspired Binarization methods for ancient documents." In *IS&T/SPIE Electronic Imaging*, pp. 72470U-72470U. International Society for Optics and Photonics, 2009.
65. Gatos, Basilios, Ioannis Pratikakis, and Stavros J. Perantonis. "Adaptive degraded document image binarization." *Pattern recognition* 39, no. 3 (2006): 317-327.
66. Su, Bolan, Shijian Lu, and Chew Lim Tan. "Robust document image binarization technique for degraded document images." *IEEE transactions on image processing* 22, no. 4 (2013): 1408-1417.
67. Feng, Meng-Ling, and Yap-Peng Tan. "Contrast adaptive binarization of low quality document images." *IEICE Electronics Express* 1, no. 16 (2004): 501-506.

68. Moro, Kamal, Mohammed Fakir, Belaid Bouikhalene, Rachid El Yachi, and Bader Dinne El Kessab. "New Approach of Feature Extraction Method Based on the Raw form and its Skeleton for Gujarati Handwritten Digits Using Neural Networks Classifier." (2014).
69. Das, M. Swamy, C. R. K. Reddy, A. Govardhan, and G. Saikrishna. "Segmentation of Overlapping Text lines, Characters in printed Telugu text document images." *International Journal of Engineering science and technology* 2, no. 11 (2010): 6606-6610.
70. Sandeep, D. R., V. B. Sandeep, and S. Dhanam Jaya. "Segmentation of Touching Hand written Telugu Characters by using Drop Fall Algorithm." (2012).
71. Pal, Umapada, A. Belaıd, and Ch Choisy. "Touching numeral segmentation using water reservoir concept." *Pattern Recognition Letters* 24, no. 1 (2003): 261-272.
72. Bharathi, J., and P. Chandrasekar Reddy. "Segmentation of Telugu Touching Conjunct Consonants Using Overlapping Bounding Boxes." *International Journal on Computer Science and Engineering* 5, no. 6 (2013): 538.
73. Saba, Tanzila, Ghazali Sulong, and Amjad Rehman. "A survey on methods and strategies on touched characters segmentation." *International Journal of Research and Reviews in Computer Science* 1, no. 2 (2010): 103-114.
74. Hastie, Trevor, Jerome Friedman, and Robert Tibshirani. "Additive Models, Trees, and Related Methods." In *The Elements of Statistical Learning*, pp. 257-298. Springer New York, 2001.
75. Loh, WeiYin, “Classification and regression trees,” in *Wiley Interdiscip. Rev.: Data Min. and Knowl. Discov.*, vol. 1, no. 1, pp. 14-23, 2011
76. Leo Breiman, “Random Forests,” *in Mach. Learn., Springer*, vol. 45, no. 1, pp. 5-32, 2001
77. Strobl, Carolin, James Malley, and Gerhard Tutz. "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." *Psychological methods* 14, no. 4 (2009): 323.
78. Yang, Mingqiang, Kidiyo Kpalma, and Joseph Ronsin. "A survey of shape feature extraction techniques." *Pattern recognition* (2008): 43-90.

79. Lu, Dengsheng, and Qihao Weng. "A survey of image classification methods and techniques for improving classification performance." *International journal of Remote sensing* 28, no. 5 (2007): 823-870.
80. Rajput, G. G., and H. B. Anita. "Handwritten Script Identification from a Bi-Script Document at Line Level using Gabor Filters." *Proc. of SCAKD* (2011): 94-101
81. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.
82. Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." In *European conference on computer vision*, pp. 404-417. Springer Berlin Heidelberg, 2006.